MILITARY OPERATIONS RESEARCH SOCIETY

INTERNATIONAL TEST AND EVALUATION ASSOCIATION

JOINT MINI-SYMPOSIUM:

HOW MUCH TESTING IS ENOUGH?

Edited by
John F. Gehrig, C. David Brown and James P. Finfera

February 28 - March 3, 1994
Williamsburg, Virginia

101 South Whiting Street • Suite 202 • Alexandria, VA 22304-3483 • (703)-751-7290 • FAX: (703)-751-8171
e-mail: rwiles@dgis.dtic.dla.mil

DISCLAIMER

This Military Operations Research Society—International Test and Evaluation Association mini-symposium report faithfully summarizes the findings of a three-day meeting of experts, users, and parties interested in the subject area. While it is not generally intended to be a comprehensive treatise on the subject, it does reflect the major concerns, insights, thoughts, and directions of the authors and discussants at the time of the mini-symposium.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	28 Feb - 3 Mar 1994	Workshop Report 2/28/94 - 3/3/94

4. TITLE AND SUBTITLE

MORS/ITEA Mini-Symposium *How Much Testing Is Enough?*

5. FUNDING NUMBERS

O & MN

6. AUTHOR(S)

John F. Gehrig, C. David Brown and James P. Finfera

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Military Operations Research Society, Inc.
101 S. Whiting Street, Suite 202
Alexandria, VA 22304-3483

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

The Deputy Under Secretary of the Army (Operations Research)
Attn: SAUS(OR)
Washington, DC 20310-0102

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unlimited; Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

These proceedings record the results of a mini-symposium held on 28 Feb - 3 Mar 1994. The goal of the mini-symposium was to provide a forum in which the military operations research and test and evaluation communities could identify key issues and develop novel and useful insights into more cost-effective test and evaluation.

The objective of the mini-symposium was to address fundamental questions that relate both to (a) the design of a test program and assessment of its use of resources and (b) the assessment and improvement of the extent to which the products of test and evaluation meet user needs. The questions addressed reflected various sources including (a) statistical committees of the National Research Council and (b) a Users' Committee that surveyed key users of T&E products and insured representation of their interests in the mini-symposium.

14. SUBJECT TERMS

15. NUMBER OF PAGES
i - xvi + 110 pages

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Sta. 239-18
298-102

CLASSIFIED BY:

DECLASSIFIED ON:

MILITARY OPERATIONS RESEARCH SOCIETY

INTERNATIONAL TEST AND EVALUATION ASSOCIATION

JOINT MINI-SYMPOSIUM:

HOW MUCH TESTING IS ENOUGH?

Edited by
John F. Gehrig, C. David Brown and James P. Finfera

February 28 - March 3, 1994
Williamsburg, Virginia

101 South Whiting Street • Suite 202 • Alexandria, VA 22304-3483 • (703)-751-7290 • FAX: (703)-751-8171
e-mail: rwiles@dgis.dtic.dla.mil

The Military Operations Research Society

The purpose of the **Military Operations Research Society (MORS)** is to enhance the quality and effectiveness of classified and unclassified military operations research. To accomplish this purpose, the Society provides media for professional exchange and peer criticism among students, theoreticians, practitioners, and users of military operations research. These media consist primarily of the traditional annual MORS symposia (classified), their published proceedings, special mini-symposia, workshops, colloquia and special purpose monographs. The forum provided by these media is directed to display the state of the art, to encourage consistent professional quality, to stimulate communication and interaction between practitioners and users, and to foster the interest and development of students of operations research. In performing its function, the Military Operations Research Society does not make or advocate official policy nor does it attempt to influence the formulation of policy. Matters discussed or statements made during the course of its symposia or printed in its publications represent the positions of the individual participants and authors and not of the Society.

The Military Operations Research Society is operated by a Board of Directors consisting of 30 members, 28 of whom are elected by vote of the Board to serve a term of four years. The persons nominated for this election are normally individuals who have attained recognition and prominence in the field of military operations research and who have demonstrated an active interest in its programs and activities. The remaining two members of the Board of Directors are the Past President who serves by right and the Executive Director who serves as a consequence of his position. A limited number of Advisory Directors are appointed from time to time, usually a for one-year term, to perform some particular function. Since a major portion of the Society's affairs is connected with classified services to military sponsors, the Society does not have a general membership in the sense that other professional societies have them. The members of MORS are the Directors, persons who have attended a MORS meeting within the past three years and Fellows of the Society (FS) who, in recognition of their unique contributions to the Society, are elected by the Board of Directors for life.

MORS is sponsored by:

- The Deputy Under Secretary of the Army (Operations Research)
- The Director, Assessment Division, Office of the Chief of Naval Operations
- The Director of Modeling, Simulation and Analysis, Deputy Chief of Staff, Plans and Operations, Headquarters, US Air Force
- The Commanding General, Marine Corps Combat Developments Command
- The Director of Force Structure, Resource and Assessment, The Joint Staff
- The Director Program Analysis and Evaluation, Office Secretary of Defense

International Test and Evaluation Association

The **International Test and Evaluation Association (ITEA)** is a not-for-profit professional organization dedicated to furthering the professional and technical interests of the test and evaluation community.

FOREWORD

The MORS/ITEA Mini-Symposium on "How Much Testing is Enough?" was an outstanding event. It was a gathering of principal decision makers and the doers of the testing, analysis and acquisition communities. It included government, industry and academia. There was a rich out pouring of insights, ideas, recommendations, and approaches in trying to optimize testing and truly understanding just how much is enough. To assure that all of this valuable information generated is not lost nor forgotten, a final report has been prepared and published. I encourage each of you to take proactive approaches to review and then implement the many sound ideas and suggestions contained within it. The mini-symposium and its final report probably generated more questions than it answered. However, I feel this is perfectly acceptable and is a good first step in the evolutionary process of optimizing and improving the acquisition/testing process.

As the Technical Chair of this mini-symposium and as the Director of the Army's Test and Evaluation Management Agency (TEMA) with responsibility for managing the Army's Testing and Evaluation budget and policy, I am taking personal interest in stimulating and monitoring the momentum and initiatives of the mini-symposium. There are several efforts that were initiated as a result of this mini-symposium or that were ongoing in parallel with this symposium. I have asked the principals of each of these efforts to keep me apprised of their progress:

a. Marion Williams, Technical Director, Air Force Operational Test and Evaluation Center (AFOTEC), and I have been tasked by the T&E Executives to chair a small tri-Service/OSD group to discuss ITEA/MORS/ADPA workshops and develop issues and recommendations for presentation to the T&E Executives. The three workshops to be reviewed are "How Much Testing is Enough?", "Emphasizing the 'E' in T&E", and "T&E Issues".

b. Under the direction of Dr. John Foulkes, Deputy Director of TEMA for Policy, the Army's Test and Evaluation (T&E) Managers addressed inadequacies of many current models and simulations. Their findings are being included as part of the RDA Domain M&S Study in support of the Deputy Chief of Staff for Operations and Plans..

c. Dr. John Foulkes is also chairing a committee composed of representatives from the Operational Evaluation Command (OEC) and AMSAA to review several selected systems to reduce testing by combining/integrating DT and OT. The final product will be a white paper to be written jointly by AMSAA and OEC to address this issue.

d. Dr. Duane Steffey of the National Research Council is a Study Director for "Statistical Methods for Testing and Evaluating Defense Systems." A final report should be out in the fall of 1996.

e. Dr. Ernest Seglie of the Office of Director, Operational Test and Evaluation is initiating a study to establish a data base of T&E Costs.

I request that other principal commands, agencies, and organizations review this report and determine if there are other issues that can be addressed. Some examples are as follows:

a. Early involvement of the T&E community in the acquisition process; involvement of the test and evaluation community in the requirements process (this process must be structured and focused on testability and evaluability).

b. Establishing a data base of identifying fielding problems that were not identified during the acquisition cycle. Possibly, the whole data base issue may be worthy of consideration. This could be a logical spin off of the Office of the Secretary of Defense Corporate Information Management (CIM) initiative.

c. Planning on how to address failures prior to actual testing (i.e. during development of testing strategies).

d. Developing-evaluation driven test plans.

e. Our need to review laws (i.e. congressional/legal language) in the spirit of reinventing the government initiatives. There may be laws that were appropriate when passed, but may no longer be applicable, or add inefficiencies to the current acquisition/testing climate. Perhaps a joint DoD and legislative task force, in a spirit of joint concerns for minimizing expenses and increasing efficiencies, should be established to address these issues.

f. Consideration of including contingency plans in the test planning process. Responses, to include the principal action officer, are requested to identify areas of interest. Many of these ideas are worth the time and effort and should generate a sound return on investment. I will assure your product will get the appropriate exposure and be presented to the Service T&E Executives. We have quality data generated from all involved with this mini-symposium. We must not lose this invaluable resource of information.

Finally, I must mention the importance of the emerging development of virtual testing as its application relates to the question of how much testing is enough. The question now becomes what is the right combination of virtual and real testing to adequately reduce risk for acquisition decisions. Unfortunately, the virtual testing concept was not mature enough at the time of this symposium to receive substantial consideration, however, its effect on the acquisition process will undoubtedly be significant and have great impact on many of the conclusions and recommendations from this mini-symposium. The application of virtual testing, its combination with real testing, and its effect on the acquisition process would probably be an excellent choice for the subject of a future MORS/ITEA mini-symposium.

I wish to express my appreciation to all who made this MORS/ITEA Mini-Symposium possible: the general chair, deputy chair, MORS, ITEA, the Panel Members, the Keynote Speaker, the Luncheon Speaker, the Invited Speakers, the Working Group Chairs and Co-chairs, Presenters,

and, of course, all the participants. We all should feel a sense of satisfaction for a job well done. However, we cannot rest on our accomplishments; we have much to do.

John F. Gehrig
Technical Chair
Washington, DC
February 15, 1995

TABLE OF CONTENTS

Chapter I

Executive Summary

Background

The question "How Much Testing is Enough?" has long plagued the acquisition community. On the surface, the answer is easy. We test to gather information to reduce the risk of applying new technology or old technology in new ways. Therefore, we have tested enough when we have enough information to reduce the risk to a level acceptable to those responsible for the application. This "easy answer" raises many very difficult to answer questions. Who is really responsible for the application and can determine the acceptable risk; the tester, the evaluator, the contractor, the program manager, the developer, the user, Congress, the media, the taxpayer, or a combination of these? Can we test until risk has been adequately reduced, or is "enough" determined by resource constraints, schedule constraints, environmental and safety concerns, a driving requirement to immediately employ the technology, or political or social considerations?

The Military Operations Research Society (MORS) and the International Test and Evaluation Association (ITEA) jointly conducted a three day mini-symposium, February 28—March 3, 1994, in Williamsburg, VA, to address the above questions. Mr. Richard Helmuth, Science Applications International Corporation (SAIC), served as General Chairperson for the mini-symposium, and Dr. Donald Greenlee, SAIC, served as Deputy General Chairperson. Mr. John Gehrig, Director, Army Test and Evaluation Management Agency (TEMA), served as the Technical Chairperson, and Dr. C. David Brown, TEMA, served as Deputy Technical Chairperson.

Goal

The goal of the mini-symposium was to provide a forum in which members of the military operations research (MOR) and test and evaluation (T&E) communities could identify key issues and develop novel and useful insights into more cost-effective test and evaluation. The objective of the mini-symposium was to address fundamental questions that relate both to the design of a test program and assessment of its use of resources and the assessment and improvement of the extent to which the products of test and evaluation meet user needs. Four working groups addressed questions reflecting various sources to include statistical committees of the National Research Council (NRC) and a user's committee that surveyed key users of T&E products and insured representation of their interests in this mini-symposium.

Approach

The approach to achieve the above goal involved first a panel discussion the evening before the mini-symposium opening to start the thoughts and discussion flowing. The next day, the mini-symposium opened with a keynote address by Dr. John Hamre, DoD Comptroller, a senior DoD decision maker who is a client of the T&E process. His ad-

dress was followed by presentations by a committee chartered to survey user needs from testing and by experts on previous related work. MG Ronald Hite, Deputy for Systems Management, ASARDA, was the luncheon speaker, providing his thoughts on the future role of T&E in the acquisition process. Four focused working groups were then formed to formulate and discuss the key issues, and a synthesis group then pulled together and documented the products of the working groups.

Discussion Areas

Generally the discussions focused on nine distinguishable areas:

(1) The impact of ATD's and ACTD's on the Testing and Evaluation (T&E) process surfaced rather frequently as an area of interest. Their relation to T&E and the subsequent effect on T&E was not well understood, and their derivatives are not sufficiently discussed in the 5000 series. T&E must quickly adapt to the changing acquisition environment, and the T&E community should take advantage of the new acquisition policies and new technologies to do its job more effectively.

(2) Over and over and throughout the symposium it was stressed that the need exists for the testers, evaluators, analysts, development test (DT) community, and operational test (OT) community to get together early in the acquisition process.

(3) There is an early requirement to organize for the pooling of information and address the how, when and what issues in the Test and Evaluation Master Plan (TEMP). Very early on, the use of prior information should be part of an iterative process that occurs within the Test Integration Working Group (TIWG). Alternative T&E concepts should be formulated and evaluated early enough to be inputs into the overall acquisition strategy of the system at the time the budget is being finalized.

(4) There was much focus on modeling and simulation (M&S) with many diverging points of view. All felt that testing and modeling are both essential, we need to pursue integrated modeling/simulation and testing, and that data gathered in the modeling and simulation process can be pooled and/or combined with actual test information. M&S activities are not necessarily either inexpensive or trivial to develop, implement, and maintain; however, the potential benefits remain significant and merit continued attention. There are few examples of distributive simulation being sucessfully applied by the T&E community. We need to look for good examples of DS for T&E because there are strong indications that testers have insufficient experience with DS.

(5) A significant portion of this mini-symposium focused on test cost. It was suggested that the true cost of testing is not clear; reducing the planners ability to streamline testing. There were many suggestions for understanding, controlling and/or reducing costs. For example, it was recommended that the cost and cost benefit of test and evaluation should be viewed from the premise that T&E is part of a program acquisition process. Also, the cost associated with a test strategy should be assessed to determine risks of a strategy. Cost/benefit analysis should be used and actual costs should be compared to estimates. An accurate data base of T&E cost/value case studies is needed. The use of prior information was potentially viable tool for reducing test cost and duration and the potential for

reducing testing costs is in combining operational testing and training (In the Navy, this is becoming the norm for the later stages of OT). However, it must be recognized that the trend is toward technologies, threats, and instrumentation that are more complex and expensive; without smarter T&E, this could potentially increase, not decrease, T&E costs.

(6) The importance of statistical application in the T&E discipline was stressed. There is a compelling need to upgrade and maintain the level of statistical interest, skills, sophistication, and appreciation in T&E. There are many statistical approaches and techniques, both standard and sophisticated, that can be utilized to increase the efficiency and economy of information collection, processing, and evaluation; however, no individual analysis tool is universally applicable or preferable in all circumstances for which it may be appropriate. T&E team members need to be adequately trained in the variety of available of statistical tools in particularly the design of experiments (DOE).

(7) Another significant portion of the seminar was directed at integration, pooling, sharing and combining of data etc. Two types of data were considered appropriate for pooling or combining: (a) specific test data from a prior phase of testing (DT or OT) of the same system; and (b) related test data from an older, but comparative system. Data gathered in the modeling and simulation process can be pooled/ combined with actual test information and vice versa. Pooling of information could occur by several methods: mingle, compare, allocate, and combine. Procedures for combining information (evaluation plans, test plans) from various testing stages (i.e. DT & OT) should be investigated.

(8) Three types of risks were identified: a) program failure, b) technology obsolescence, and c) life & limb. T&E should attempt to characterize uncertainties & risks. The question is "how?".

(9) Even though many have the misconception that the issue of laws and regulations is an unapproachable topic, the discussion was reasonably free and candid. There is a general feeling that the time has come to re-look at the existing laws and regulations. Some expressed that the promulgation of regulation, policy and procedures over the years has been inconsistent and misinterpreted. There was a recommendation that legislation on flexibility in funding between testing and procurement was needed to allow for addressing the unexpected.

Conclusions and Recomendations

Particular attention should be directed toward the conclusions and recommendations generated by the Synthesis Group. They generated three sets of recommendations that warrant serious consideration:

- The T&E planning process should be subjected to fundamental Operations Research (OR) scrutiny, including explicit consideration of alternative T&E strategies, contingency planning, and cost / benefit tradeoffs.

- Each individual T&E program should empower a small, stable, "integrated T&E team" (with some mix of contractor, developer, trainer, operational test and evaluation (OT&E), user, and office of the secretary of defense (OSD) representation) to - manage design, evaluation, and implementation issues; continually monitor T&E planning activities and review emerging

results; and revise t&e plans as warranted. The team's dual emphasis should be on comprehensiveness and efficiency.

- More emphasis should be placed on ensuring that each individual test and evaluation activity is efficiently designed and analyzed. In particular, experimental design techniques and other established statistical approaches should be better exploited.

Even though there was no single answer for "How much testing is enough?" there were many positive constructive recommendations for optimally reducing, improving, and streamlining testing. There also were insights into existing and potential problems and concerns which will play significantly into optimizing all aspects of testing. This mini-symposium has laid significant ground work for the effective growth and evolution of the defense testing process.

Chapter II

General Sessions Presentations

A. Panel Discussion

General

The panel discussion was held on the evening before the opening of the mini-symposium. The purpose of the panel discussion was two fold. First, because most of the symposium attendees were from the test and evaluation (T&E) community, the panel started the symposium with some input from the rest of the acquisition community. To assure this input, the invited panel members were high level individuals representing all services who generally believe that the T&E community continually attempts to dictate that "too much" testing is conducted. Second, panel members were chosen that were not afraid to speak their minds, and held opposing views with each other and with the symposium attendees. The panel members represented the Army, Navy, and Air Force, and did indeed stimulate a discussion that provided an excellent beginning to a very productive mini-symposium.

Panel Members

The panel members were: Mr. John Gehrig, Moderator, LTG William Forster, Military Deputy to the Assistant Secretary of the Army (Research Development and Acquisition); RADM George Strohsahl, Commander, Naval Air Warfare Center; Mr. James O'Bryon, Deputy Director, Test and Evaluation, Land and Maritime Systems; Office of the Under Secretary of Defense (Acquisition and Technology); Mr. John Gilligan, Air Force Program Executive Office, Combat Support Systems; Mr. George Williams, Army Program Executive Office, Tactical Missiles. No attempt was made to report the verbatim comments of the panelists. The following are the key points made by each.

Presentations

LTG Forster. LTG Forster stated that the question of how much testing is enough is a function of how early in the development cycle testing is accomplished. The earlier it is accomplished, the less testing is required and the greater the value. He also stated that we need to have the users involved as early as possible and that modeling and simulation are replacing prototyping. He charged the audience to test to learn, not to fail, and to not test just to comply. In fact, he said we should do away with all laws requiring testing. Finally, LTG Forster stated that he has concern that the independent evaluator has license to scope the test without fiscal responsibility for support of this scope of testing.

RADM Strohsahl. RADM Strohsahl proposed renaming the mini-symposium to "How much risk are we willing to accept?" He proposed testing only to the highest level of risk acceptable to the customer. He also predicted that commercial off-the-shelf acquisition and adoption of commercial procedures may possibly extend the T&E process. One reason is the loss of process control provided by military specifications and standards. RADM Strohsahl recommended that more testing be integrated (not applied serially) by contractors and Government testers and by

both DT and OT. He said that we need an integrated system test environment which includes models and simulations. Finally, he urged all to apply the lessons of TQM to the test process, i.e., listen to the testers in the field as to how to do it better.

Mr. O'Bryon. Mr. O'Bryon stated that there are many reasons to test. These include reducing program risk, to meet statutory requirements, to add discipline to the development process, to reduce costs, to save lives and equipment, to better understand complex processes, to push the frontiers of science and technology, and to participate in the scientific method. However, he stressed that testing without resulting action accomplishes little. He stressed that neither testing nor modeling by themselves, no matter how much, will suffice. They are both essential. Some suggestions of how to get the necessary test data but save time and money are to: piggyback some currently planned DT, merge testing and training where appropriate, assure that the purposes of testing are clearly articulated prior to testing, develop more efficient data and information archival and dissemination systems to allow maximum value from accomplished testing, and focus on major operational requirements and not on the minutia. In discussing risks, Mr. O'Bryon identified three types: risk of program failure, risk of technological obsolescence and risk of life and limb. Testing must be conscious of all of these risks. In response to a question asking what legislative changes each panelist would like to see, Mr. O'Bryon indicated that he would like to see legislation which would allow some flexibility in funding between testing and procurement to enable unexpected problems arising out of testing to be addressed without major disruption to programs. In response to another question, he suggested the concept of small "built-in holds" in the procurement schedule, comparable to the space launch countdowns, to allow opportunities for assessing and correcting program deficiencies without major impacts on program viability.

Mr. Gilligan. Mr. Gilligan said that information systems characterized by frequent releases of incremental improvements are not served well by our current OT&E process which takes too long and is too expensive for these systems. Testing needs to be done in a "system of systems" context, but that can be very risky because of the impact of failure on the component systems. He cautioned all to be wary of simplistic metrics such as "software maturity."

Mr. Williams. Mr. Williams proposed that the right amount of testing: demonstrates the weapon system performance, provides sufficient data for modeling and simulation verification, answers the technical issues, allows for successful operational assessment, reduces program risk, and enhances probability of cost effective production. Some actions to pursue are to: attempt to cut costs by testing and evaluating smarter, share information both vertically and horizontally, improve team spirit among all DoD participants such that all share ownership of the developing system, utilize power-down decision making (thus reducing oversight), strive to develop a seamless test environment, include the materiel developers as an integral part of all test planning efforts, actively pursue test methodologies for cost reduction, create hybrid models for testing and simulation, pursue integrated modeling/ simulation and testing, and support active test instrumentation development to include funding.

Discussion. Several panel members addressed a question about the impact of ATD's on the T&E process. The first point made was that the DoD 5000 is totally inadequate when it comes to ATD's. The second point is that rigor in the testing of ATD's (as opposed to "experimenting") could be instrumental in the saving of future testing required.

The panel was asked to define "risk". Program managers can always identify where the risk is, but cannot define how to recognize a level of acceptable risk. Usually the decision of "acceptable risk" is defined in terms of cost, however, the risk that we will accept in a program should depend on the consequences of failure.

B. Keynote Address - Dr. John Hamre DoD Comptroller

Dr. Hamre set the tone for the mini-symposium with his views on T&E. He stated upfront that his views came mainly from his prior vantage point on the staff of the US Senate and not so much from his present position as DoD comptroller.

Dr. Hamre described how DOT&E was created during the Cold War climate envisioning major systems requiring large operational tests to support the decision to go into full rate of production. He stated that this politics of the 1980's demanded independent Operational Testing ("capital O, capital T"), but the reality of the 1990's is that production rates are smaller and schedules are less demanding and therefore operational testing's role should change as a result ("lower case o, lower case t"). His belief is that operational testing should not have its present status as statutory testing outside of the acquisition process, but should be imbedded within the acquisition process just as developmental testing is now. He questioned whether the large elaborate structure of OT&E makes sense, but emphasizes the value that the rigor of OT&E has brought to the acquisition process.

Addressing the balance between testing and modeling and simulation, Dr. Hamre stated that testing should be done just enough to calibrate some M&S upon which the decisions will be based. He said, "We are entering an era where simulation is the only way we can really test some of the interesting systems."

C. Users Survey Report on "How Much Testing Is Enough?" - James Duff, Technical Director, COMOPTEVFOR

In support of the MORS/ITEA mini-symposium on "How Much Testing Is Enough?," a committee of leading DoD T&E experts conducted a study to determine the needs of T&E Data Users. The study consisted of three phases (defining the users, review of T&E governing directives and a survey of the users' needs).

Results of the user definition indicated that there are many users. Examples of the users surveyed included; War fighter, DoD/Service Decision Maker, Program Manager/Sponsor, Manufacturer, R&D centers and Taxpayer.

Results of reviewing the T&E governing directives indicated that responsibility for determining "How much testing is enough" is not defined in any of the governing directives and therefore the tester/evaluator becomes responsible by default. The only possible exception noted was that for operational testing. The Director, Operational Test and Evaluation, by his charter, determines the number of items required for low rate initial

production and he has oversight for all operational test and evaluation.

One hundred and two T&E Data Users were surveyed relative to three key questions:

- What do users need from test & evaluation?
- What will they do with what they ask for?
- What needs are not currently being adequately satisfied?

The following encapsulates survey findings:

- The "Customer" is not well defined. "The eye of the beholder" determines one's needs as War fighter, DoD/Service Decision Maker, Program Manager/Sponsor, Manufacturer, etc.
- Test objectives/operational requirements are not well defined.
- Lack of confidence and the "unknowns" drive the T&E process. Modeling and Simulation viewed with skepticism. System interaction with environment the only true test.
- DT and OT management is in consistent..players not on "the same sheet of music". TEMP requirements should be locked in to ensure continuous validity of test results.
- Testing should be conducted to determine the limitations & bounds of performance, not to validate minimum thresholds only.
- Users focus on test deficiency correction, less on overall system performance/ successes.
- Test reports are too voluminous..."All things to all customers" but..Distance was also a factor...The farther from Washington, the higher the satisfaction with the T&E product.

User Survey results were presented to the T&E Mini-Symposium members at the opening general session to provide a "perspective framework" for follow on discussion.

D. Summary of National Research Council Presentation to MORS/ITEA Symposium on How Much Testing is Enough? - Dr. Duane Steffey, National Research Council

The talk was structured in three parts: (1) a brief introduction to the National Research Council, (2) a report on the Workshop on Statistical Issues in Defense Analysis and Testing, and (3) a progress report on development of a multi year panel study of statistical methods for testing and evaluating defense systems.

The National Research Council (NRC) is the principal operating agency of the National Academy of Science and National Academy of Engineering. In this capacity, the NRC advises the federal government and provides services to the public and the scientific and engineering communities. The NRC has a long working relationship with the Department of Defense. Operating units within the NRC that have worked extensively on military projects include the Air Force Studies Board, the Board on Army Science and Technology, and the Naval Studies Board.

At the request of the Department of Defense, the NRC's Committee on National Statistics, in conjunction with the Committee on Applied and Theoretical Statistics, held a workshop in September 1992 on statistical modeling, simulation, and operational testing of weapon systems. Defense analysts were invited to write and present background papers and discuss substantive areas in which they sought improvements through application of

statistical methods. Statisticians and other participants responded by suggesting alternative approaches to specific problems and identifying problem areas that might especially benefit from the application of improved statistical methods.

Several major themes emerged from the workshop. Because testing is expensive and potentially dangerous, it is important that tests be designed using statistical principles of experimental design to permit the efficient collection and analysis of test data. Informed decision making requires understanding all sources of variability in an analysis. More formal attention to analysis of sensitivity to model assumptions, validation of models, sampling and non-sampling source of errors, and selection biases would contribute to this goal.

The analyst's responsibility in presenting results is to ensure that the uncertainties from an analysis are reported to decision makers. Use of graphical methods may assist in presenting quantitative information in a way that avoids technical jargon and makes policy implications clear. Statistical methods of combining information and borrowing strength across experiments could be employed to use information from earlier stages in designing and analyzing operational tests.

The classical approach to statistical hypothesis testing is problematic, because the asymmetry of significance tests leads to unproductive arguments about what the null hypothesis should be, and, hence, where the "burden of proof" lies in the testing process. More neutral approaches based on statistical decision theory would provide a more appropriate conceptual frame work.

Test efficiency and effectiveness could benefit from moving toward a more neutral and cooperative environment in managing quality in weapon systems with emphasis on achieving consistent improvement rather than clearing interim hurdles at program milestones. Data are a precious resource in the Defense Department and could be used more effectively. Creating new capability in data archiving and management—e.g., developing a rational data base—coupled with a statistical and data analysis unit could improve the DoD's use of data.

Copies of the report *Statistical Issues in Defense Analysis and Testing: Summary of a Workshop* can be obtained by calling (202) 334-2240 or write to: Committee on National Statistics, National Research Council, 2101 Constitution Avenue NW, Washington, D. C. 20418.

The Committee on National Statistics is developing a multi year panel study on statistical methods for testing and evaluating defense systems. The objective of the study is to improve the effectiveness and efficiency of testing and evaluating defense systems as part of the defense acquisition process. The panel will include expertise in statistical, operations research, software engineering, military systems, and defense acquisition. In its work, the panel will consider the measures of operational effectiveness, the structural design of operational tests, methods to incorporate information from previous analyses, and ways to present uncertainties in test results.

The panel is to hold its initial planning meeting in March 1994. The proposed schedule calls for an interim report in the spring of 1995 and a final report in the summer of 1996. The panel intends to approach its work by

conducting retrospective case studies of past and current systems. In this manner, the panel will gain access to the test data and personnel knowledgeable about particular systems and will gain an appreciation of the environment in which operational tests are conducted.

The panel goals are to learn from the test and evaluation community and to subsequently make a scientific contribution to defense testing. Symposium participants were invited to offer comments and suggestions for the study, particularly with regard to selection criteria and candidate systems for case studies.

E. Identifying Research Needs and Problem-solving Tools for Test and Evaluation - Donald P. Gaver, Professor Of Operations Research, Naval Postgraduate School

Summary: Research Directions For Increased Test and Evaluation Effectiveness

- Anticipate and quantify cost-effectiveness of testing options (field, modeling and simulation, and hybrid).
- Recognize *all* costs: testing facilities, life-cycle cost of system, including operational deficiencies, environmental impact, training requirements,...
- Characterize uncertainties and risks (not all random/narrowly statistical). Attempt to prevent, but anticipate, and provide to adapt to *surprises*.
- Archive and *use* data: history, field test, operational experience. Invest in retrospective data analysis and development of institutional memory. Anticipate and develop ways of dealing with and compensating for incomplete, missing, generally messy, data.
- Focus on understanding the *operational contribution* of new (or upgraded) systems.

Classification of Systems

Items to be tested range from complete systems to subsystems to components; from upgrades and modifications to new starts. Various technologies are in use or proposed, and costs and operational environments and utilities also vary widely and are uncertain.

An attempt to *classify* system types, and to compare prospective new systems to others of its type could sharpen the test and evaluation process. The comparison should also include alternative ways of meeting threats and accomplishing missions.

Models and Simulation

Modeling and simulation (M&S) is a purposeful collection of assumptions ("facts") that can lead to efficient answers to important questions concerning proposed or actual new systems. The advantage: (relatively) easy (numerical) responses to questions; the disadvantage: degree and dimensions of trustworthiness and communicability of the answers.

There are several formal M&S types and styles: *physical* (based on similar physical or "real" situations, e.g. drones or other imitation targets); *mathematical* (symbolic, computer-activated); *hybrid* or combinations of these.

Informal models, (e.g. sets of assumptions combined by experience alone) are always used when planning and managing test and evaluation. Formal M&S should be used to supplement and strengthen, not replace, the informal. Formal models are always wrong to

some degree, but this is an advantage as well as a flaw. Formal models can serve to efficiently archive institutional memory, and to promote communication and highlight and resolve uncertainties, and to bring sharper focus on issues. *In particular* the development of formal models that project the military effectiveness of new systems in new environments should be energetically pursued.

Statistics in Test and Evaluation

The recently published National Research Council/National Academy of Science publication *Statistical Issues in Defense Analysis and Testing: Summary of a Workshop*, authored by John Rolph and Duane Steffey, provides a critical appraisal of the situation with some suggestions for improvements. The remarks below generally supplement, and are meant to add some emphasis to, the message of the above document.

Acquisition of field data could become more meaningful if, prior to actual test, outcomes were first simulated and these results analyzed. The value of the new information provided by (expensive) field test could be assessed in advance, and the field test design adjusted to provide needed information efficiently. The first purpose of statistics is to add information by *reducing uncertainty*.

Standard textbook-style statistical procedures for data analysis (especially classical hypothesis testing or acceptance sampling using traditional but arbitrary α and β errors) have been typical in the community, as has uncritical "sequentializing" of such procedures. Tests have been treated in isolation, without use of prior information. This situation can be improved by using decision analysis that explicitly incorporates costs and operational contributions. Legitimate procedures for combining information from various testing stages, for instance developmental testing (DT) and operational testing (OT) should be investigated; this will involve care in adjusting for systematic bias. Research in this area is currently underway at the Naval Postgraduate School.

Use of regression models and especially *response surface methodology* seems natural in the operational testing environment. For instance, response could be probability of detection by range r of an incoming threat (missile, aircraft, etc.); explanatory variables could be its altitude, speed, type, plus weather conditions. Learning effects could be quantitatively represented. Non-linear and discrete-response regression models are now available (as above) and can be brought to bear. Suitable textbooks are by Box, Hunter and Hunter (1978), (fractional factorials and response surfaces); McCullagh and Nelder (1984); Cox (1974), (binary regression).

Computer packages for these applications exist. It is desirable that they be used with appropriate background and training, plus common sense.

Example

Cases are useful for conveying the flavor of the testing activity. They highlight the conflict between testing adequate to reveal problems and the economy and possible military advantage of prompt fielding. Dr. E. Seglie of Defense Operational Test and Evaluation, has provided hypothetical case studies that illustrate situations that may be encountered in practice. One such concerned testing of an upgraded (not entirely new) surface-to-surface tactical missile. The previous (be-

fore-upgrade) system had been tested to the extent of around 12 missile shots so as to check reliability and damage expectancy. The "new" missile experienced changes in motor, guidance, range warhead, launch software, targeting, and targets.

The requirement was that the upgrade be an improvement (how much?). Suppose distributed interactive simulation shows "no difference". Despite such an outcome, often there now remains pressure to go ahead with a field test, perhaps (because it is an upgrade) concentrated on a selected small part of the region of the new missile responsibility. Such a proposal presumably reflects distrust of the simulation model, or its application to the current situation. Certainly a careful review of the model's capability is called for to help resolve the issue. In part this is an investment in the future, presuming that the model will be used again.

In the present hypothetical case the model results were overridden and tests made: missiles were fired and records of fragmentation products used to calculate a damage expectancy on a target. The (estimated) damage expectancy so obtained fell below requirements and the upgrade threatened with failure.

In such circumstances, there may well be an attempt to invoke statistical data-analysis alternatives so as to argue that the system should pass. In the present case it was proposed to use a *median* of calculated damage assessments rather than the usual mean; summary by the median produced a number that technically exceeded the threshold. The mean apparently responded to some extremely low values that dragged *its* number below the threshold. An attempt to justify this suggestion was apparently made on the basis of "statistical robustness". For discussion see Mosteller and Tukey (1977).

Statistically robust procedures are often legitimately invoked to reduce the effect of gross or outlying measurement errors; development and use of such procedures has enjoyed deserved popularity but the implications of the numerical calculations need to be interpreted with care. Such measurement error was not the issue for the missile: the low values that affected the mean were structural, resulting from too-frequent misperformance of a system element. The fact that these values occurred and affected the mean *should* sensitize the system developers to a system problem to be rectified: properly interpreted, the median-mean comparison acts as a *diagnostic* that triggers a technical correction. The moral or lesson may be that use of alternative and novel statistical concepts is neither to be discouraged nor uncritically embraced, but that the results require careful interpretation and suitable action.

Bibliography

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer-Verlag, New York.

Box, E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters*, John Wiley, New York.

Marshall, K. T. and Oliver, R. M. *Decision Making and Forecasting: with Emphasis on Model Building and Policy Analysis*, Forthcoming McGraw-Hill book.

McCullagh, P. and Nelder, J. A. (1984) *Generalized Linear Models*, Chapman and Hall, New York.

Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*, Addison-Wesley Publishing Co., Reading, MA.

Oliver, R. M. and Smith, J. Q. (ed.) (1990) *Influence Diagrams, Belief Nets and Decision Analysis*, John Wiley, New York.

Chapter III

Working Group Deliberations

A. Working Group I - Cost/Benefit Consideration of Test Programs

Chairperson:

Mr Joe Rech
Resource Manager
USAF Test & Evaluation, Resources Division

Co-chairs:

Mr Ed Eagar
Washington DC Operations
Georgia Tech Research Institute

Dr Gerry McNichols
President
Management Consulting and Research, Inc

Mr George Wauer
DOT&E
OSD

Dr Al Benton
USA Materiel Systems Analysis Activity

Introduction

Working Group I's specific focus for "Cost/Benefit Consideration of Test Programs" was:

1) How should the cost and benefits of tests and test programs be evaluated?

2) What metrics should be used for measuring the value of T&E, both prior to testing and after fielding?

3) Since T&E is an open-ended process, how should the risks of test termination be assessed?

4) What is the balance between decision risk and test costs?

5) How should the scope of development testing be traded off with operational testing?

6) What is the value of test data to estimate logistic support requirements?

Approach

To generate background information and present several initial viewpoints, two briefings and four formal presentations were provided to the entire assembled working group.

Briefing and Presenter

"A T&E Process," Mr Joe Rech

"T&E Requirements," Mr Ed Eagar

Presentation and Presenter

"A Proposed Method for Bounding System Performance When Testing is Cost Prohibitive", G. S. Schwartz, Capt, USMC

When tests evaluating system performance through success frequencies are very costly, small sample plans must often be chosen as a matter of necessity. There are suggested methods for using sequential and zero defect

sampling to bound system performance as an alternative to gross assumptions of homogeneity of multifactorial effects. Knowledge of similar systems and likely operational scenarios provides a basis for establishing bounds and rules for determining if performance thresholds have been met.

"Current Issues in Live-Fire Planning, Analysis, and Testing," Messrs J. Terrence Klopcic and William E. Baker

Investigation of various methodologies which would quantify the costs, risks, and benefits associated with full-up, live-fire testing and with various alternative testing strategies. Such methodologies should serve as a formal procedure for the request of a waiver from full-up, live-fire testing of a system. Four methodologies proposed; all incorporate considerable subjectivity and vary considerably in the total effort needed for implementation. There are tradeoffs between simplicity, thoroughness, and perceived accuracy. Valid arguments exist in favor of each proposal. A single final cost/risk/benefit methodology may evolve from some hybrid of the proposed methodologies.

"Cost of Testing Versus Program Costs," Ms Peg A. Mion

Historically, attempts to determine the percentage of program costs expended on test and evaluation have begun with research of the costs of successfully completed programs. However, there are many problems associated with determining and understanding the cost of testing for major programs, and this information is not always easy to obtain. The Army Tactical Missile System (TACMS) was chosen for this research effort because: 1) It had a well planned T&E strategy, 2) It is a recent major program, 3) It completed development in a relatively short period of time. A methodology was offered for obtaining test costs. Additional research regarding the cost of developmental testing (including production and post-production testing) was done on four other major programs which confirmed the developmental test cost ratios developed for the Army TACMS.

"A Simple Decision Aid for Determining Initial Test Size," Messrs William D. Moore and Anthony Zimmermann

During initial planning for large scale operational tests of complex expensive intelligence systems, it is necessary to balance the cost of the test and the amount of data required to obtain statistically valid results. This planning can exert a profound effect upon the cost of a test. For example, in the case of a unit fielding two expensive systems (one per platoon), does it make economic sense to test two systems (platoons) or can valid data be obtained testing just one system? In essence, can valid results be obtained at less cost? For Measures of Performance (MOP) which are stated as pass/fail for a given threshold, the test can be considered as a series of Bernoulli trials which is described by the Binomial Distribution. If the number of such trials is known for both proposed test sizes, then exact Binomial confidence intervals for each MOP can be easily calculated. Given the upper and lower limits for each case, the size of the interval is easily obtained. If one assumes that the size of the confidence interval for a given sample size is a crude measure of the validity of the expected results, then by comparing the sizes of the confidence intervals, it can be seen if it is really necessary to assume the extra cost of increased testing.

Following the briefings/presentations, the group was split into two working sessions. The task of Group 1A was to answer questions 1, 5 and 6 (Introduction, above); and, the task for Group 1B to answer questions 2, 3 and 4.

QUESTION 1: **How Should Cost And Benefits of Tests And Test Programs Be Evaluated?**

There should be an evaluation process:

1. Understand the acquisition strategy.
2. Develop alternative T&E strategies.
3. Develop the assessment criteria - cost/benefit/risk.
4. Estimate potential consequences of each strategy relative to the criteria.
5. Develop an assessment of alternatives.
6. Refine.

The cost and cost benefit of test and evaluation should be viewed from the basic premise that T&E is part of a program's acquisition process. Inherent therein is the acquisition strategy for a program. For each T&E strategy, an assessment criteria is developed based on cost/benefit versus risk along with an estimate of the potential consequences of each strategy.

Assessment Criteria:

Benefits:

- Traceable to Program Requirements (user needs).
- Utility of Information Learned.
- Responsive to Program Planning (COEA).
- Addresses the 'Unknown Unknowns'.

Costs:

- Ratio of T&E to Program Costs.
- Other Ratios.
- Success Oriented Budgeting.
- Cost/Benefit Analysis.
- Actual Costs vs Estimate.
- Test Process Compatibility.

The criteria should assess the test strategy and planning relative to program requirements (user needs) to ensure traceability within the planned test methodology. The test strategy should assess the utility of information to be derived to confirm that it contributes to the decision process and is responsive to program planning documentation such as the COEA. To the extent possible, strategies should be assessed for the ability to anticipate the unexpected consequences (unknown unknowns). Criteria for cost versus risk may be expressed in several ratios such as DT&E$:Program$, DT&E$:OT&E$, or any number of other relationships. The cost associated with a test strategy should be assessed to determine risks of a strategy if success oriented budgeting is used. Wherever possible cost/benefit analysis should be used and actual costs should be compared to estimates. Lastly, the test strategy alternatives should be assessed relative to compatibility with the test process to increase early test knowledge (such as modeling/ simulation) that contributes to future testing at other facilities such as hardware-in-the-loop or ultimately open air range testing.

QUESTION 2: **What Metrics Should Be Used For Measuring The Value of T&E, Both Prior to Testing And After Fielding?**

The underlying issues that arose during the presentations and working group discussions were the need for historical case studies and a

consistent cost methodology. An accurate data base of T&E cost/value case studies is needed. Many have started this process on their own, looking at one or a few programs in detail, but a full-fledged, fully supported effort is needed. The T&E community has to agree on a work breakdown structure for test costs. Standard assumptions need to be made when it comes to constant dollars, program cost and life cycle costs. Definitions must be consistent across the community: What is a test? What is an experiment? Also, infrastructure capability and cost (i.e., fixed versus variable) need to be defined.

The metrics for test and evaluation benefit should address "confidence in knowledge" about the system and determine if the system is mature enough to progress into the next stage. The amount and quality of knowledge needed is dependent on factors such as: past testing, future testing, and maturity of the system. Testing addresses all three areas of cost, schedule, performance. Normal 'risk management' is structured against known areas of concern and tests can be structured to address these. But, risk management must also include the identification of "unknown unknowns" as a desired benefit of testing. Some metrics in this area may relate to "safety factors" and "how much of the operating envelope has been covered". After fielding, testing can verify solutions to problems or shortfalls identified during OT&E or a P3I in progress. The bottom line is the need to compare apples to apples (test costs to test costs) and have a data base of case studies.

QUESTION 3: **Since T&E Is an Open-ended Process, How Should The Risk of Test Termination Be Assessed?**

Termination of a test (not due to the loss of funds) could occur under the following circumstances:

- If terminated due to failure, consideration should be given to the potential value of lost data (e.g., How much is further data needed now? How valid will further data be, given the failure? How much will it cost to obtain data late?)
- If production testing is being accomplished, normal or previously accepted QC standards may allow early termination.
- Termination due to success and exceeding performance expectations is possible, and positive.

Truncation (due to loss of funds) requires some major 'planning ahead'. The primary emphasis needs to be on identifying the high priority information to be obtained through testing and a plan to achieve this data gathering as early as system maturity allows. The bottom line of test termination is lost confidence in the knowledge desired to make informed decisions.

QUESTION 4: **What Is The Balance Between Decision Risk And Test Costs?**

Decision risk, or the level of confidence and knowledge gained as a result of testing (see Question 2), is inversely related to test costs; however, it is non-linear. As test costs climb the decision risk lowers asymptotically and the decision maker must weigh the 'added value' of greatly increased costs versus small decreases in risk. Decision risk is difficult to measure and the criteria varies by program, ACAT, risk aversion of the decision-maker, and so forth.

QUESTION 5: **How Show The Scope of Developmental Testing Be Traded Off With Operational Testing?**

The scope of DT versus OT was considered to be less important than the ability to share or possibly reuse the data for multiple purposes. What was considered of high importance is the desire that test planning be developed from a 'team approach' leading to increased potential for shared data between test phases and synergies of test objectives relative to DT versus OT. Recent initiatives directed at obtaining common instrumentation should be leveraged to yield increased standardization of data, resulting in reuse or possible transportability of the data. To accomplish this reuse of data, initiatives should be directed at methods to archive and certify the data so that future retrieval is meaningful and credible.

QUESTION 6: **What Is The Value of Test Data to Estimate Logistic Support Requirements?**

Similar to the preceding question, rather than segregate test data, the emphasis is to be directed at test design that would lead to data that would be usable for multiple purposes. Proper test design should also yield only that test data with utility...unused 'by-products' should not exist. The test data outputs through out the acquisition process should contribute to logistic support considerations (i.e., provisioning, maintainability, MTBF, product improvement, etc).

Summary

The cost and cost benefit of test should be viewed from the basic premise that T&E is part of the acquisition process. Inherent therein is an acquisition strategy. Subsequently, a test strategy should assess the utility of information to be derived, to confirm that it contributes to the decision process and is responsive to program planning. Such strategy should anticipate the unexpected consequences (unknown unknowns). An accurate data base of T&E cost/value case studies is needed. Standard assumptions need to be made when it comes to constant dollars, program cost and life cycle costs. The metrics for test and evaluation benefit should address "confidence in knowledge" about the system and determining if the system is mature enough to progress into the next stage. In the area of "unknown unknowns" metrics may relate to 'safety factors' and 'how much of the operating envelope has been covered'. Risk of test termination is a factor of: cause, further data need/validity/cost, and priority of information. Decision risk is difficult to measure. The criteria varies by program, ACAT, risk aversion of the decision-maker, and so forth. The scope of developmental versus operational test is less important than instituting a 'team approach'—leading to increased potential for shared data between test phases and synergies of test objectives relative to DT vs OT. Initiatives should be directed at methods to archive and certify data so future retrieval is meaningful and credible. Proper test design leads to data usable for multiple purposes; yields only test products/data with utility.

B. Working Group II - Optimization of Test Programs

Chairperson:

Raymond G. Pollard III
Technical Director
U.S. Army Test & Evaluation Command

Co-chairs:

Mr. Brian Barr
Research Staff Member
Institute for Defense Analyses

Ms. Christine Fossett
Assistant Director
Government Accounting Office
Office of Policy

Dr. Donald Gaver
Professor of Operations Research
Department of Operations Research, Naval Postgraduate School

Within the broad context of "How Much Testing is Enough" is the question of "How do I get exactly enough?" In short, optimization of the test program. Working Group II had two excellent presentations. The first, by Dr. Jim Streilein, AMSAA, and Dr. Al Benton, AMSAA, presented a comprehensive review of options and solutions in optimizing test and evaluation. The second by Dr. Jim Elele, USAEPG, presented an excellent approach for the use of statistical design to optimize testing to a PM's cost bogey. (Dr. Elele's report is appended in its entirety).

The problem statement Working Group II was asked to address is straightforward. Simply put--can we optimize test programs and how?

Second, we should strive to minimize the program cost to obtain that critical information. Linking the two objective functions is the area of risk assessment. The extremes are to obtain all the information you can from testing with cost not a constraint and reduce the risk, hopefully to zero. Note that this does not necessarily imply a credible decision.

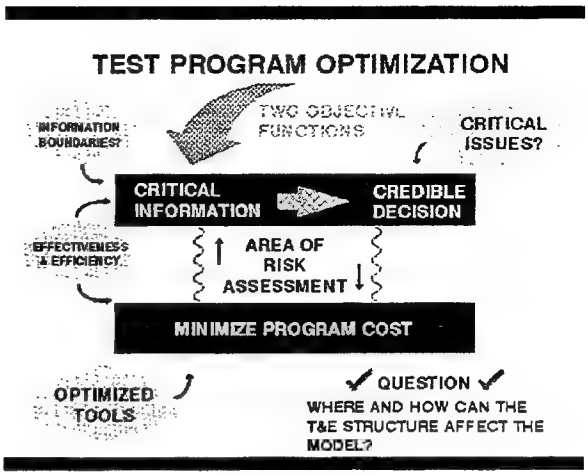

Figure 1

It was helpful for Working Group II to build a reasonably simple model to define test program optimization (Figure 1). The model suggests there are fundamentally two objective functions. First, we must obtain the critical information to reach a credible decision. Second, we should strive to minimize the program cost to obtain that critical information. Linking the two objective functions is the area of risk assessment. The extremes are to obtain all the information you can from testing with cost not aconstraint and reduce the risk, hopefully to zero. Note that this does not necessarily imply a credible decision, only that the information is there to reduce the risk of the unknown and, hopefully, provide the basis for good analysis. On the other hand, if we invest nothing in testing, and presuming no other source of information, we have a "data free" analysis and a decision that assumes maximum risk, all else being equal. The point is not that testing per se reduces risk, but that any trade-off or substitution of information should have a corresponding risk assessment.

Around the edges of the basic model we note that a credible decision, and therefore the information to support that decision must be keyed to the critical issues for that decision. We note also, in the accumulation or search for critical information, that boundaries exist to the flow of information. There were noted instances of development testers (DT) not releasing information to operational test (OT) evaluators, of OT evaluators determining they can not accept "tainted" DT information, of government not accepting contractor test data, etc.

Effectiveness and efficiency are goals: effectiveness tends to drive objective function 1, e.g., you want the test to be sufficiently effective so as to obtain that required information; and efficiency drives objective function 2, e.g., minimize resource expenditure.

It is important to recognize that optimization of testing must be attacked at three distinct but interacting levels. Individual tests must be optimally designed to do no more repetitions than are necessary to answer the test objectives. A key point is to recognize that, at a level below test program optimization is the development and use of optimized tools to support both efficiency and effectiveness.

Each system must then have a coordinated optimized program that allows individual tests to provide data to address more than one issue or objective. Thus, for example, a developmental test could provide feeder data to both the Live Fire evaluation and to the operational evaluation while still meeting its primary DT objective.

Finally, there must be coordination between programs to develop overall test programs that optimize the use of test assets between test programs for more than one system.

The question then is where and how can the T/E "structure" affect the model?

The first issue Working Group II tackled was the question of whether there are measures and procedures that can be implemented to improve the optimization of test programs. The belief is that there are at least two key areas where we can do better. First is data sharing. We must break down the boundaries before we test to ensure that we know whether or not there is information that obviates the need to test. If the critical information can be obtained at lesser cost than testing, or better yet if it is available, then we have gone to the heart of objective function 2. We have minimized cost to the program. Having determined the information set that is available, then we key to the critical evaluation issues and the "blank spaces" in our data set and focus on the operative question, "What do I need to know that I don't know and that I can only obtain from a physical test?"

The above simple steps (which require some not so simple thinking) should, if fully executed, minimize the information required from test and therefore go straight to objective function 2 in reducing program cost. While the process seems appealing, and in some degree is used, it should be explicitly included in all evaluation plans. At a minimum it will raise the consciousness of the evaluator, at best it will result in substantially less testing.

The second issue asks how you include failure loops in test strategies. The easy answer is you just do it. It is, however, must easier said than done. Most programs, partic-

ularly the major programs, are typically success oriented, schedule (time and money) driven to meet budget cycles, IOC's, etc. We have known for years that, in the development process, we should be event driven—it's not going to happen! So what do we do?

Frequently, perhaps typically, we develop the test plan by asking ourselves what is the time frame, what are the dollar constraints and what are the decision points. How do we determine the minimum amount of information to support a decision. To the testers' and evaluators' credit they typically call out the risks and assumptions associated with the sporting course laid out. Also typically, the risks and assumptions become increasingly less visible as the program competes for funding. When failure happens, programs panic, they restructure, they allocate new dollars and defend the system against the growing perception at senior levels that the "program is in trouble."

To alleviate that, we recommend that, up front, we plan on how we will address failure. We should identify the high priority, high technical risk areas as we currently do. However, instead of simply assigning risk in a success oriented plan, we should take the next step and plan for potential failure. This could take the form of several alternatives. For example: we may wish to invest a little more up front in parts or systems; we may wish to modularize test programs and structure resources such that downstream tests can be automatically moved in to slots created by failures; we may wish to have "on-call" contracts for software trouble shooting experts who are, at the front end, read into both test and system software.

Contingency planning should be a necessary part of every plan from TEMP down. For a program that has no problems and meets its success oriented schedule, the test contingency costs will be an added cost but, without execution, a manageable burden. For a program that experiences failure at some point it should save cost. In any event it will provide a defense to the perception of a "program in trouble" if the service can articulate that what happened was both "not unexpected" and planned for.

Can test programs be adapted to emerging results? Yes they can but there is some risk associated with adaptation of test programs in the perception that we are fitting the requirement to the result. Haphazard or spur of the moment adaptation opens the door to that perception even if the adaptation is entirely appropriate and well meant. So how do you plan for adaptation?

We should design evaluations and tests for adaptation. The easiest way is to do the a priori analysis to define where you can truncate tests when you know enough for a credible decision—objective function 1. But, even if you cannot foresee specific results, good analysis may suggest that, given we know the nature of the expected data, we can consider a set of possible outcomes that lead to changes in test designs. For example, we may need to expand a test when we need to know more than we originally thought or to alter a test when we need to know something different than we originally thought. The fact that it is hard to think in the "contingency" world does not detract from the benefit of so doing.

Where data volume is such that rapid (near real time) collection, reduction and analysis is not possible or practical, we should explore some techniques to get at the essential mes-

sage of the data at an ordered level such that decisions can be made to stop or change testing early. We should investigate tools to do this, e.g., response surface analysis is a possibility. If we are either incapable or, under the press of work, not available to do this internal to DOD, we should seek high level outside scientific scrutiny and assistance perhaps at the National Science Foundation "level of capability."

Is disciplined, flexible testing possible? We believe it is. By example, the highly structured and disciplined live fire test process also has inherent flexibility. As events occur and changes happen, the test structure is established to permit very rapid decisions on options. The fact that the decision level is very high in the structure does not mean that flexibility is sacrificed for the overall test program. A less resource committed, somewhat less structured and lower level decision authority can also reach a disciplined but flexible level. What must be done is to ensure that:

- test plans are evaluation driven to key on the critical data
- decision making is taken to the lowest level consistent with priorities (and note that the lowest appropriate level may be OSD or it may be the test director on the ground).

The thrust should be down

- that contingency planning is used
- that *a priori* analysis to obtain pretest prediction of results (as in live fire testing) is employed and,
- that, whenever possible, embedded instrumentation is employed to reduce variability and "unique solutions."

All of this requires the testers and evaluators to be involved early in programs, during program design and technology development.

But—it won't just happen. The current situation is haphazard. There is no inherent process by which the T&E community is in early. Where testers and evaluators know of technologies and programs at inception, and where they have approached technologists and program managers and where they have been accepted, the knowledge transfer to technologists and developers in terms of testability in design occurs. Similarly the knowledge transfer to testers and evaluators in terms of system parameters and system design have permitted early preparation for the ability to test and for innovative T&E design. This is similarly true where technologists have sought the support of testers' capability and evaluators' analytical capability and where developers and program managers have sought the testers and evaluators as part of the concurrent engineering team.

To achieve full effect of both discipline and flexibility, testers and evaluators must, inherently and institutionally, be part of the early acquisition process. This may be particularly important as the acquisition process evolves to ATD's and ACTD's.

Are there better ways to state testable requirements? Yes. We have recognized in the past that a tester and evaluator review of requirements (and specifications, RFP's, etc.) is of benefit. However, the coordinating process associated with hundreds of systems dilutes the focus. To optimize the use of testers and evaluators, they should, as a matter of course, be part of the combat developer's requirements working groups for ACAT I systems. Their role should not be to tell the

user what the requirements should be but rather to define the testability or at least the evaluability of requirements. If requirements must appropriately be written beyond the ability to test or at least evaluate (and some would say they should not) then at least ensure that the user/combat developer has a clear understanding of that early in the program. Early in the program they should articulate the fact that the requirement can be neither tested nor evaluated and not be in a position of being branded a failure in the end because it cannot be tested.

From a program cost standpoint, elimination of the requirement may save considerably. At least, by eliminating the requirement, capital investment to attempt to test the requirement may be substantially reduced.

Are there techniques for minimizing testing? Yes and they are employed in the T&E community every day. Examples include combining DT and OT; "piggy-backing" live fire tests on DT or OT; testing two or more systems together (for example, lead the fleet testing with add-ons for other programs at USA Aviation Technical Test Center); statistical design techniques; use of models and simulations; pretest predictions; and Test Optimization and Quality Analysis Method (TOAQAM) as mentioned in Dr. Elele's paper. Most of these are being done already, but not in all programs. The test and evaluation community and the program mangers must keep pushing!

Considerable anecdotal evidence was presented within the working group to suggest that the processes which link T&E, acquisition, requirements generations and the analytical base for those requirements are:

- flawed
- misunderstood
- unwieldy
- unresponsive
- inconsistent
  "
  "
  "
- "pick a word"

This is not to say that the overall process is broken or hopeless. It is not even to say that any particular process is beyond repair. We are also focusing here not on the DOD 5000 series nor the umbrella process. Nor are we focusing on the SOP's and test procedures typically used at both development and operational test facilities.

The focus is in that middle ground between the OSD policy and the on-ground technician. It's in the arena where agencies must interact, where the team approach in joint working groups, in test working groups and in tieing it all together is essential. Services do it differently among themselves and often within themselves. The promulgation of regulations, policy and procedures over the years has been inconsistent, misinterpreted and, in many cases, had a deleterious effect on "getting the job done." This is compounded by the experience of the players who, despite entreaties to "break the mold" and be innovative, are nonetheless too often trapped by their experience—the way it has always been done or the way they think it is supposed to be done or, in fact, the way they have to do it.

There is ample evidence that the Test and Evaluation Master Plans (TEMP) and the Test Integration Working Groups (TIWG) that are responsible for developing them are not effective tools for optimizing test programs. The

focus is too often on individual words and sentences, or on single tests designs, or on resourcing tests. Rarely is there real substantive work done to design a well coordinated and optimal overall test program focused on answering the agreed upon set of critical issues.

This working group recommends that a substantive, constructive, non-confrontational, independent team conduct an in-depth (several month) look at the "middle ground" to include the TEMP TIWG process. The approach should be systemic looking at all levels of the process to determine not only what is happening, but also why it is happening. It should explore both the internal and external forces that cause the system to be the way it so that the root causes of the problems can be identified and eliminated.

Those selected for this team should be able to bridge the gap between a broad perspective of how the process currently works at the integration level but also how it works where the rubber meets the road. They should be inclined to help rather than criticize and should, preferably, have a reasonable level of organizational independence (one suggestion was a select set of knowledgeable retirees).

In summary, test program optimization is achievable and is currently sought (with considerable success) in all services. But, substantially more can be achieved through early involvement, good analytical planning and the use of modern (and perhaps not so modern) tools. Focus on the specifics of tools may be enhanced by the application of and outside scientific help to the problem. Finally a constructive and substantive process review by an independent team appears to be beneficial.

Working Group 2 studied the optimization of test programs. Measures and procedures that can be implemented to improve the optimization of test programs include data sharing (are there information or procedures that obviate the need to test?) and a priori analysis (What do I need to know that I can get only from a test?). Including failure loops in test programs requires up front planning on how to address failure to include identifying high priority, high technical risk areas for contingency planning. Test programs can be adapted to emerging results but there is always a risk that this will be perceived as fitting requirements to results.

Working Group 2 felt that disciplined, flexible testing is possible by creating evaluation driven test plans, planning for test contingencies, using pretest predictions of results, and having embedded instrumentation. Most importantly, techniques for minimizing testing include combining DT and OT, piggybacking live fire tests on DT and OT, testing systems together, using statistical design techniques, and using models and simulations, a priori to designing tests, to replicate tests and to extend test results.

Appendix 1. ***Concepts for Efficient/Reduced Test and Evaluation*** **- Dr. James Streilein, U.S. Army Materiel System Analysis Activity**

With the known and expected drawdowns in defense over the next few years, it was considered imperative that the Army make the most efficient and effective use of resources in developing new systems and for upgrading fielded systems for the future. At a January 1992 conference, the Deputy Under Secretary of the Army for Operations Research presented a number of ideas aimed at reducing

testing and making the process more efficient. The DUSA(OR) presented several "Old" and "New" concepts. They included:

- Single integrated test plan—How much contractor testing can be credited?
- Field experiments to examine tactics, training, etc prior to IOTE
- Early on use of SIMNET
- Quit when you have learned all you can learn.
- Combined DT/EUT
- Multiple items in single test. Pros? Cons?
- Events within trial as sample—not trials.
- Sequential testing approaches.
- Interval test estimation versus hypothesis testing.
- COEA in support of test alternative selection.

At the next conference, many additional ideas, concepts and examples were provided by the test and evaluation community for expanding and implementing his ideas.

AMSAA consolidated a list of plausible ideas from the Army Test and Evaluation community for reducing test associated costs and making the test and evaluation process more efficient. The effort was initiated as a result of DOD's cost reduction emphasis further advocated by the Army Acquisition Executive and the Deputy Under Secretary of the Army for Operations Research. Most of the proposed concepts are not new and are being implemented within the community. All of the ideas may not be applicable to every system or program, but the concepts are provided for consideration and renewed emphasis.

Each concept is first assigned to a test strategy, test alternative or test/item specific category. The test strategy concepts are those related to general policy or scope of test and evaluation. Those identified as test alternatives represent test program management alternatives for peculiar items/ commodity areas. Finally, the test/item specific concepts are applicable to a specific test of a specific item. All of the concepts and their categories are identified below:

- Test Strategy Concepts
  - Single Integrated Test Plan
  - Establish Appropriate Requirements
  - Field Experiments
  - Reliability Emphasis
  - Involve Tester/Evaluator Earlier
  - Testing Oriented To Customer
  - Core Resources
  - Full Tech Base Participation
  - Expand Soldier's Role During Development
  - Reliability Physic-of-Failure Design
- Test Alternatives
  - Maximize Contractor Efforts/Data
  - Joint Development/Technical Test and Operational/User Test
  - Multiple Items Per Test
  - Cost and Operational Effectiveness Analysis on Test Alternatives
  - Modeling/Simulation
- Test/Item Specific
  - Results Indicate Test Termination
  - Test Design/Sample Size
  - Sequential Testing
  - Interval Versus Hypothesis Test
  - Combine/Piggyback Subtests
  - Surrogates
  - Enhanced Test Instrumentation
  - Learn From Previous Test Results

- Increase Decision Risks
- Limit System Test/Retest
- Maximize Use of Drivers

More specific information about each concept was then presented. Each of the concepts identified above are first related to each acquisition phase where it may apply. Then, possible proposals to address/implement each concept are summarized along with benefits, concerns and/or limitations that may apply. An example for some of the concepts is also provided.

Minimizing test costs has been a major thrust for many years. Many of the concepts are being applied during the test design and Test Integration Working Group process, but some may warrant renewed Senior Level emphasis. The future's tightening budget climate will dictate continued application and increased emphasis on test cost reduction, accepting greater risk and the application of new cost reduction ideas.

Appendix 2. ***Tester's Choice: To Field Test or Not*** **- Dr. James N. Elele, United States Army Electronic Proving Ground, Fort Huachuca, Arizona**

Abstract

Testers are frequently required to determine the performances of various military systems under realistic operating environments. This requirement often demands that field tests be performed in order to determine the capability of the system under test (SUT). In some instances, the demand for field testing is placed just for the feeling of confidence it brings. While field testing is very desirable and essential in many cases, oftentimes, the variability of test parameters under field environments, together with the fact that humans are prone to making errors, make field testing results of questionable value relative to the high cost associated with conducting these tests. Thus, there is a need to determine the conditions under which results of field testing can be too expensive, or may not provide the necessary information required to determine the performance of the SUT. This paper develops an approach for solving this problem using statistics.

In developing the methodology, we assume that there is a band of performance levels, F(p) (with $a \leq F(p) \leq b$, for $p_0 \leq p \leq p_1$), required of the SUT. The range of values the parameter, p, can assume determines the range of values F(p) can assume. The value of p is measured in a field test and can vary randomly or systematically depending on the field environment. We then try to answer the following question: To what level must we limit the variability of the parameter, p, to determine whether the SUT met the condition that F(p) must remain in the specified performance band? If p varies wildly under field environments, a large number of tests would be necessary to meet the performance criteria, and thus field testing would be very expensive. Alternatively, if the variation in p is such that F(p) does not move a lot from the specified performance band, then field testing the SUT would be reasonable.

Introduction

As the military continues to downsize, the high cost associated with testing systems in the field demands that only the most critical items be tested in this manner. Even in these critical cases, the austere fiscal environment requires that field tests be conducted only in situations where the results provide valuable information that could not be obtained through any other

testing approach. When field testing is found to be inappropriate, other methods must be used to determine the ability of the system under test (SUT) to perform as expected before being committed to actual operation. It may happen that field testing is the only appropriate test approach for a given SUT. In that case, the system must be field tested. However, we lack an established objective way to show with confidence that the sought-after SUT information cannot be obtained more cheaply by an alternative test approach.

To assist decision makers in determining which of the various SUTs are suitable for the different types of testing (i.e., field testing, hardware-in-the-loop/laboratory testing, modeling and simulation, or a combination thereof), objective quantitative methods must be developed for assessing the suitability of different test methodologies, and the value of the information they provide relative to cost. In essence, we must develop test quality control methodologies, coupled with test optimization schemes, for use in the allocation of the scarce resources available for testing. In this paper, personnel from the United States Army Electronic Proving Ground (USAEPG) use organizational experience from long-time involvement in testing to develop such a test quality control/ optimization methodology.

Our approach is to combine well-established statistical and optimization methods and our existing Integrated Test Methodology (ITM) for the design of test procedures that provide the needed information while allocating available resources in an efficient way. The ITM combines modeling and simulation, hardware-in-the-loop/laboratory testing, and field testing, as appropriate, for a complete evaluation of the SUT. The new test optimization and quality analysis method (TOAQAM) we are developing starts by using statistics and available information (such as manufacturers' data, equipment specification data, in some cases data from preliminary testing, and data from existing models and simulations) to estimate the number of measurements needed to establish the performance required of the SUT. From the number of measurements required, we can obtain a reliable estimate of the cost of the test. By asking whether the variability of the parameters being measured allows useful information to be obtained from the measurements in a cost-effective way, we are able to select/combine alternative test approaches that best provide the required information in the most efficient manner.

Subsequent sections present the ITM and the simple statistical methods used to estimate the number of measurements required to establish the performance of an SUT and the optimization approach used for resource allocation. Then we show how these are used together for TOAQAM. We do not consider it appropriate to go into high-level theoretical analysis here because this may be a hindrance for people who are interested in everyday use and the practical application of the methodology. The methods we present were selected for their simplicity and ease of use in everyday, real-life applications. However, we shall indicate possible directions for future and more advanced theoretical work on TOAQAM.

The Integrated Test Methodology

USAEPG, at Fort Huachuca, Arizona, is chartered with performing electronic equipment testing of command, control, communications, computers, and intelligence (C^4I) systems in support of the US Army Test and Evaluation Command's (TECOM's) mission. To support this mission, USAEPG has devel-

oped the ITM. In this methodology (see Figure 2), modeling and simulation, hardware-in-the-loop testing, and field testing are used to complement each other for a comprehensive evaluation of a system being tested. ITM has been used to evaluate such systems as the Single Channel Ground and Airborne Radio System (SINCGARS), the Joint Tactical Information Distribution System (JTIDS-2M) terminals, and the Mobile Subscriber Equipment (MSE) system.

To support the ITM for electromagnetic environmental (EME) testing, we employ USAEPG's large accumulation of EME data, various systems models, laboratory and hardware-in-the-loop test facilities (e.g., the Stress Loading Facility), as well as the vast instrumented range facilities for field and outdoor testing. The TOAQAM approach

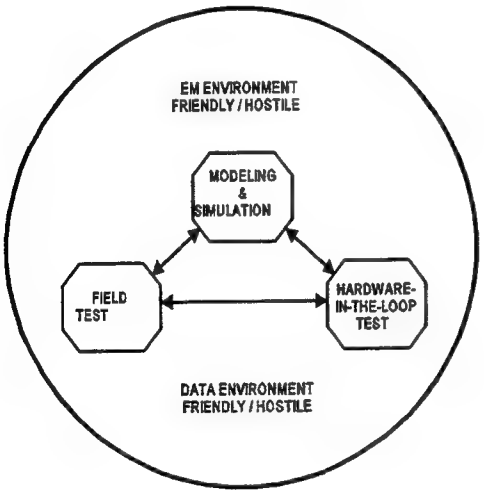

Figure 2. Integrated Test Methodology

supplements the ITM with statistical analysis of available data (manufacturer's information, test data from similar systems, published and unpublished literature surveys, data from existing models and simulations, etc.). From the result of this analysis, we determine and follow an optimum path of the ITM, using constraints on available test resources as a guide to minimize cost.

Statistical Methods for Estimating Number of Test Measurements

Our goal here is to use statistical analysis to predict conditions under which conducting a test can be too expensive, or may not provide valuable information for evaluating the performance of the SUT. Our premise is the well-known statistical principle that *the larger the variability in a population, the larger the number of samples required to obtain a "good" estimate of the attributes of that population.* From this principle, we show that at least for systems with constrained performance requirements, if the variability consistently exceeds a certain maximum value, only minimum valuable information is obtained unless the number of measurements is large. In a field test situation, for example, we can control the variability of only a minimum number of the measured parameters. Consequently, the test resource optimization technique, TOAQAM, requires that if the influence of the uncontrolled variables (such as those imposed on the test by the physical environment) figures prominently in determining the performance of the SUT, one must then ask whether field testing is still the most efficient option.

Most military testing is done to determine the performance/effectiveness of the SUT, and to train and acquaint the soldier in the use of the SUT. For performance/effectiveness evaluation, the SUT is tested to specified criteria, usually referred to as the Required Operating Capability (ROC) for operational tests. The ROC is usually specified for a given battlefield

scenario or environment. For example, the ROC for a C^4I system can be stated for a specified jamming-to-signal ratio (e.g., a certain message completion rate can be specified under a certain jamming environment). Thus, the specified condition for the ROC determines the critical parameter of interest, p, to be measured in a test in order to determine the performance of the SUT. The range of values assumed by the critical parameter of interest is determined by the set of variables, $V = (v_1, v_2, ..., v_n)$, that one tries to measure through testing. Therefore, the level of variability in the elements of V determines the level of variability in p and, subsequently, the level of variability in the performance of the SUT under the specified test conditions.

If F(p) represents a measure of performance of the SUT as specified by the ROC, for example, then the facts stated in the last paragraph imply that F(p) depends on the critical parameter, p, in some way, and that the critical parameter, p, depends on the set of variables, V. Thus, any test performed on the SUT can be regarded, in an abstract way, as sampling from the elements of the sample space, V, for the purpose of drawing inferences on the properties of F(p). This abstraction allows us to apply statistical methods to test and evaluation, and provides us with a means for drawing some inferences regarding the requirements for the actual testing before test start.

The Structures of Performance/Effectiveness Specifications

It is interesting that in all practical applications, the performance/effectiveness requirements specified for any SUT can be classified into one of two conditions. These two conditions are statements of the type: (1) $F(p) \geq F_{min}(p)$, or (2) $F_{max}(p) \geq F(p) \geq F_{min}(p)$. It turns out that the type 1 condition is usually more common than the type 2 condition. However, the type 2 condition is mathematically much more general because the type 1 condition is a subclass of the type 2 condition. We shall elaborate on these conditions in what follows.

Condition 1: $F(p) \geq F_{min}(p)$

This type of specification requires that the SUT's performance under specified conditions must be greater than or equal to a certain minimum. Thus, this type of ROC specification divides the performance space into two mutually exclusive regions: one in which performance is acceptable, and the other in which performance is unacceptable, with the dividing line being the specified minimum performance, $F_{min}(p)$. This is illustrated in Figure 3.

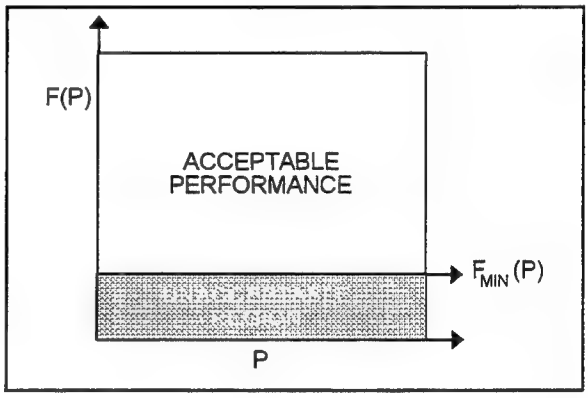

Figure 3. Structure of Performance Specification, Condition 1

Number of Measurements Required for Performance Determination under Condition 1

When performance is specified as in Condition 1 above, the aim of testing is to determine with a specified level of confidence if a specified proportion of the values of F(p) will

fall above $F_{min}(p)$. The specified proportion of the values of F(p) falling above $F_{min}(p)$ usually means that the SUT is "passing the test." A larger than specified proportion of F(p) falling below $F_{min}(p)$ usually means that the SUT is "failing the test." Thus, an interesting question to ask from a test optimization point of view is: How many measurements are needed to determine with α-percent confidence that β percent of the values of F(p) will fall above $F_{min}(p)$? In practical terms, this is the same as asking how much testing does one need to determine (through performance) whether the SUT is passing or failing. If a large number of measurements is required, the cost can be high depending on the unit cost of each individual measurement.

Answering this critical and seemingly difficult question is simple if one approaches the problem statistically. To illustrate, notice that if F(p) is specified as in Condition 1 then for statistical purposes F(p) can be considered a binomial random variable, since it can only assume two values (pass or fail) for each measurement. In such a case, the number of measurements required to estimate F(p) to within an error bound of, say d for example, from its true value, with a confidence of $(1-\alpha)100$ percent is given by

$$n = \frac{Z_{\alpha/2}^2}{4d^2}$$

where $Z_{\alpha/2}$ is the standard normal variate that leaves an area of $\alpha/2$ to the right.

Under these circumstances, the probability that F(p) will assume any specified value (in the passing or the failing region) based on the estimated number of measurements can be computed from the binomial distribution.

Alternatively, we can assume that F(p), the performance being determined, is a normally distributed random variable. If this is not the case, the correct distribution can be determined before proceeding, or use can be made of well-known distribution-free statistical methods. After this choice is made, the following prescription can be used to determine n, the number of measurements required:

1. Assume a distribution for F(p) (we have assumed normal distribution for this illustration), or use distribution-free statistics.

2. Select the confidence level, α, and the proportion, β percent of F(p) required to fall above the minimum performance, $F_{min}(p)$, for the SUT to be passing.

3. Using available data or using data from preliminary test results (in the absence of any data, start by approximating with a randomly generated sample obtained from the assumed distribution), estimate the value of the mean performance, $\bar{F}(p)$, as well as the performance standard deviation, σ_F.

4. Using a one-sided Statistical Tolerance table (STT) (e.g., see Natrella's experimental statistics handbook in the bibliography), find the value of n, the sample size, for which the following equality holds:

$$F_{min}(p) = \bar{F}(p) - \sigma_F K$$

where the factor, K, is determined from the one-sided (normal) STT in the following way:

   a. Guess the value of the number of measurements, n.

b. Using this value of n and the selected values of α and β, determine the value of the corresponding K from the STT.

c. Using the obtained value of K, check whether the above equality holds.

d. If the equality holds, the value of n is the correct number of . . . measurements required, then stop. Otherwise, go to the next step.

e. Repeat steps a through d.

5. If a (normal) STT is not available, one can compute K from the following equations:

$$U = 1 - \frac{Z_\alpha^2}{2(n-1)}$$

$$W = Z_\beta^2 - \frac{Z_\alpha^2}{n}$$

$$K = \frac{Z_\beta + \sqrt{Z_\beta^2 - UW}}{U}$$

where Z_α and Z_β are standard normal variables, and are obtained from standard normal table.

Condition 2: $F_{max}(p) \geq F(p) \geq F_{min}(p)$

This type of performance/effectiveness specification requires that the SUT's performance stay between two specified limiting (performance) values under specified conditions. This specification divides the performance space into three distinct regions: a lower region in which performance is unacceptable, a middle region in which performance is acceptable, and a top and final region in which performance is also unacceptable. This type of performance specification is much more general than the first type. In fact, the type specified in Condition 1 is a subset of this one, since we can obtain the Condition 1-type specification from this one by simply requiring that $F_{max}(p)$ be located at infinity. The Condition 2-type of performance specification is illustrated in figure 4.

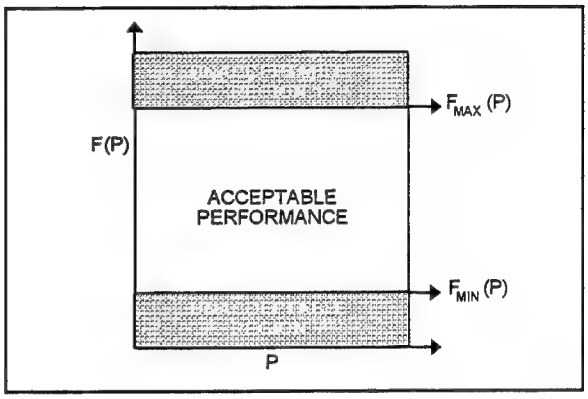

Figure 4. Structure of Performance Specification, Condition 2

Maximum Variation Allowed to Keep F(p) in the Specified Performance Band under Condition 2

Larger variability in a population translates into the requirement for a larger number of samples for obtaining "good" estimates of the attributes of that population. We must seek, therefore, to determine *a priori* the limiting variability level in the parameter p (as determined by the variabilities in the measured set of variables V) required to restrict the performance measure, F(p), in the specified region of acceptance. If variability in p consistently pushes F(p) away from the acceptance region of performance, then either the equipment is failing the test because of SUT design or operational failure or no useful information is

being obtained from the test. In this case, an inordinate amount of measurements will also be needed to get a good estimate.

To derive the required bound, we observe that this problem is similar to the statistical control problem in which we wish to find out how much variability can be allowed in the input to a process to keep the output within specified control limits. The difference here is that we are not at liberty to control the inputs, rather we take what the physics and the environment give us. It is the fact that we cannot control these variations that allows us to conclude that continuing testing if these variations exceed a certain limit does not provide valuable information relative to the cost of the measurements.

To obtain the needed bound, suppose we define F(p) as a normal independent, random variable (if it is not, we can repeat the analysis using the correct distribution, or by means of distribution-free statistics) required to take values in a variable performance band of average diameter μ_w. We let σ_w represent the standard deviation of the band diameter. Now, if we have another independent, identically distributed normal random variable, $F_F(p)$ (representing the field-measured performance), with mean μ_F and standard deviation σ_F, and we wish to control $F_F(p)$ in such a way as to keep $(1-\alpha)100$ percent of its values within the specified average window width, we may then ask: What is the maximum allowable variation in $F_F(p)$ that keeps $(1-\alpha)100$ percent of its value in this window? To answer this question, we define a new random variable, Y, by

$$Y = F_w(p) - F_F(p) \qquad (1)$$

where the subscript F indicates any values assumed by performance measured in the field, and the subscript w indicates values of performance falling inside the specified performance band. Observe that if $Y < O$, then $F_F(p)$ is outside the band. Since Y is obtained by a linear combination of variables, we may take the expectation of Y to obtain

$$\mu_y = \mu_w - \mu_F \qquad (2)$$

Now the standard deviation of Y is given by

$$\sigma_y = \sqrt{\sigma_w^2 + \sigma_F^2} \qquad (3)$$

Since $F_F(p)$ and $F_w(p)$ are random normal variables, then the new variable defined by

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \qquad (4)$$

is a standard normal variate. We are interested in the probability that Y is less than zero $P(Y \leq 0)$, but we know that

$$(Y \leq 0) = P\left(Z \leq \frac{0 - \mu_Y}{\sigma_Y} \right) \qquad (5)$$

The requirement that $(1-\alpha)100$ percent of $F_F(p)$ fall within the mean diameter of $F_w(p)$ then translates into

$$P\left(Z_o \leq \frac{0 - \mu_Y}{\sigma_Y} \right) = \alpha \qquad (6)$$

where Z_0 is the value to the standard normal variable Z, that leaves an area of size α to the left. Substituting (3) into (6), we obtain

$$P\left(Z_o \leq \frac{-\mu_Y}{\sqrt{\sigma_w^2 + \sigma_F^2}} \right) = \alpha \qquad (7)$$

Also, substituting for μ_y from (2), we obtained

$$P\left(Z_o \le \frac{\mu_F - \mu_w}{\sqrt{\sigma_w^2 + \sigma_F^2}}\right) = \alpha \qquad (8)$$

For any specified value of α (or rather(1-α)), we can obtain the value of Z_0 from a standard normal table. For example, if α = 0.05, Z_0 = -1.645, and if α = 0.01, Z_0 = -2.335. From equation (8), we observe that if we want (1-α)100 percent of $F_F(p)$ in the specified performance band, then

$$Z_o \le \frac{\mu_F - \mu_w}{\sqrt{\sigma_w^2 + \sigma_F^2}} \qquad (9)$$

Solving (9), we obtain

$$\sigma_F \le \sqrt{\frac{1}{Z_o^2}\left(\mu_F^2 - 2\mu_F\mu_w + \mu_w^2\right) - \sigma_w^2} \qquad (10)$$

Equation (10) places a bound on the level of variation allowed in the field measured performance, $F_F(p)$, if (1-α)100 percent of its values must fall inside the required window of performance defined for the SUT. For practical purposes, this means that if the variation of the measured values of performance, as determined by its standard deviation, exceeds the value specified by the right-hand side of the last equation, then most of the test results will result in information outside the performance window of interest. Hence, depending on what performance means for the SUT, either the system may be failing or the test is providing information in an area of no interest to the tester.

Number of Measurements Required to Estimate the Variability of Performance within a Stated Precision

In the last section we derived a bound for the level of variation allowed in the field-measured performance, $F_f(p)$, if (1-α) 100 percent of its values must fall inside the required window of performance. Such a bound has practical value, especially if we are interested in estimating the number of measurements needed to establish the performance of the SUT within the stated performance bounds. The number of measurements required to estimate the variability of F(p) (as determined by its standard deviation) within a stated percentage of its true value has direct correlation with the actual total cost of the measurements. For example, in designing preliminary testing, one would want to know the number of preliminary measurements required to obtain acceptable estimates of the parameters of the SUT. This information would be needed for the design of the main test.

The number of measurements required to establish the variability in performance, σ_F, within α percent of its true value with confidence β is equal to the "degree of freedom plus one." Using α and β, the degree of freedom can be obtained from published statistical tables.

Number of Measurements Required to Estimate the Average Performance of the SUT within a Stated Precision

In various test situations, we are only interested in determining the average performance of the SUT under specified conditions. In such circumstances, test design and test cost estimation require that the number of measurements necessary for the determination of average performance be determined a priori. By assuming that performance, F(p), is

approximately normally distributed, we can estimate n, the total number of measurements required to determine the average performance to within d-units of its true value from the following:

$$n = \frac{t_{(1-\alpha/2)} S^2}{d}$$

provided we are willing to accept the risk that we will be wrong α percent of the time. Thus, α has to be chosen small. Here, S^2 is an estimate of the variability of performance (see above discussions), and $t_{(1-\alpha/2)}$ is the value of the t-statistic that leaves an area of $\alpha/2$ to the right tail of the t-distribution, and can be obtained from a statistical table.

Notice that to get as close as possible to the true value of average performance (i.e., to make d as small as possible) from a highly variable measurement (i.e., large S^2 value) requires a large number of measurements (i.e., increases n). From this observation, and because most variations induced by field environments are usually uncontrollable, one must consider the viability of all possible testing alternatives before committing totally to field testing in this period of reduced funding.

Testing under TOAQAM

The TOAQAM testing scheme superimposes pretest statistical data analysis, the application of quality control design methods, and the use of mathematical optimization on the well-established ITM developed by USAEPG. The result is the optimum use of limited test resources for obtaining maximum information on the SUT, while minimizing the cost of the test. A top-level illustration of this methodology is given in Figure 5, followed by a simplified example of test optimization.

A Simplified Illustration of Test Optimization under TOAQAM

In the TOAQAM process, circumstances may exist where the optimum test scheme consists of juxtaposing field test, hardware-in-the-loop test, and simulation and modeling in a complementary way for complete evaluation of the SUT. This is primarily the ITM, except that constrained optimization and statistical design are used to determine how much of each of the three testing methods is used. This process can become very complex in real-life situations. However, by doing this work in advance, the tester can be sure that the resulting test scheme is the best that the available funds can provide.

To illustrate the approach, we consider a test consisting of taking repeated measurements a number of times. Using the result from these measurements, the tester is supposed to determine whether the system satisfies required performance capability of the type specified by Condition 1 or Condition 2. Suppose that all the measurements could be made in the field, in the laboratory, or by simulation. Given the choice and enough funds, the program manager (PM) would prefer field testing. However, because of funding limitations, the PM wants to use a combination of simulation, hardware-in-the-loop testing, and field testing (field testing to give the PM some confidence in the conclusions). The question is, in what proportion must we allocate the measurements to each of the three test methods to make optimum use of the limited funding and to determine the SUT's performance?

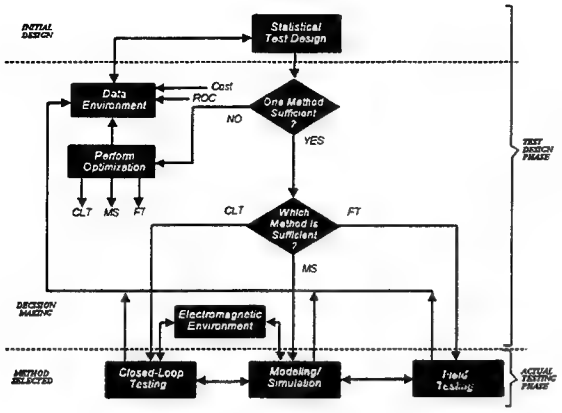

Figure 5. Test Optimization and Quality Analysis Method

Mathematically, this translates to the following optimization problem:

$$\text{minimize } [C_T] = \sum_{i=1}^{3} c_i n_i = \text{Total cost of test}$$

Subject to:

1. $\sum_{i=1}^{3} n_i = n$ = Overall total number of measurements

   i.e., sum of the individual types of measurements, n_i (for I=1=simulation, I=2=hardware-in-the-loop test, I=3=field test), must equal the total number of measurements, n, as determined statistically from the performance specification.

2. $\sum_{i=1}^{3} c_i n_i \le C$ = Total funds provided for the test

   i.e., total cost of the three types of tests must equal the fund provided, where c_i is the cost per measurement for test type I; n_i is the total number of measurements of type I; and C is the total funds provided for the test.

3. $F_{max}(p) \ge F(p) \ge F_{min}(p)$ or $F(p) \ge F_{min}(p)$

   This is the performance constraint used to determine n from statistical analysis. Typical estimation procedures for n have been included in this paper.

4. $$c_i n_i \ge 0, \; i = 1,2$$

   representing total cost of simulation and hardware-in-the-loop testing, respectively. These costs must be zero or positive, since either some or none of these types of testing must be done.

5. $$c_3 n_3 > 0$$

   representing the cost of field testing. It must be positive to ensure that some field testing is done.

Solving this problem gives the proportion of the total number of measurements to be allocated to the different types of test methods to satisfy performance requirements within the specified minimum cost. While this is a much simplified example (and a real-life problem will be much more difficult to formulate), it illustrates how much one can do when careful consideration is given to the use of available mathematical techniques in design, analysis, and test optimization. A sample solution for this simplified test optimization is given below.

Problem: Determine the number of measurements of each type. Notice that there are many possibilities. The one chosen depends on the PM's preference. But at least we have offered him/her a choice.

Test Method	Cost per Run ($)	Number of Runs		
Simulation and Modeling	90	140	5	5
Hardware-in-the-Loop Testing	100	32	194	182
Field Test	150	28	1	3

Total No. of Measurements, n = 200
Total Budget = $ 20,000
Cost per Unit Measurement:

Simulation	=	$90/Measurement
Hardware-in-the-Loop Test	=	$100/ Measurement
Field test	=	$150/Measurement

Figure 6. Sample Optimization Run

Summary and Conclusion

This paper presents the evolving test optimization and TOAQAM scheme currently being developed by USAEPG government personnel. The TOAQAM initiative originated from the reality that the ongoing reduction in size of the military requires new approaches for military testing and test designs. The knowledge base accumulated at USAEPG from over 30 years of military testing experience provides a unique asset in developing new testing schema. Consequently, the TOAQAM scheme is based on USAEPG's already existing ITM.

The TOAQAM scheme superimposes statistical analysis and optimization on the well-established ITM. Thus, the new scheme uses statistical estimation to determine the number of measurements to be made in a test, based on the structure of the required performance of the SUT, and the level of variability of the critical parameters and variables being measured in the test. The test is then optimizedby minimizing total cost, subject to the constraints of available funding, the proportions of measurements to be allocated, as appropriate, to the different types of test and evaluation methods (i.e., modeling and simulation, hardware-in-the-loop or laboratory testing, and field testing), and the specified performance requirements. The result is that the tester is not automatically locked into one form of testing or the other, whether it is the most appropriate or not, and maximum information is obtained on the performance of the SUT at a minimal cost.

Using the statistical methods presented in this paper, one can show both quantitatively and qualitatively, whether field testing would be an economical alternative, as well as whether the variability of (field) measured parameters would allow useful conclusions to be drawn from the measurements. This allows the tester to decide whether to seek alternative test approaches or continue with field testing. When one particular type of testing or evaluation is not the only viable approach, optimization methods, such as that illustrated in the simplified example presented in the paper, could be used to guide decision makers on how to select the optimum proportion of

measurements to be made by the three types of test/evaluation, to obtain maximum information on the performance of the SUT.

The methods presented here are simple, powerful, and well-established; however, more work remains to be done. For example, there is a need to extend the analysis on the influence of variability on the number of measurements and performance to the case where the sources of variation are identified, the variation decomposed to its components, and the variation minimized. This is what analysis of variance (ANOVA) and experimental design are developed for, and the need exists to extend this type of work to these very widely used techniques.

Bibliography

Anderson, T.W., *An Introduction to Multivariate Statistics,* John Wiley and Sons, Inc., New York, 1985.

Bronson, Richard, *Theory and Problems of Operations Research,* Schaum's Outline Series, McGraw Hill Publishing Company, New York, 1982.

Douglas, Johnny A., "Threat Simulators," 35th Annual JEWI Symposium, Colorado Springs, Colorado, March 1990.

Elele, James N., and James L. Cole, "The Concept of Realistic Battlefield Environment (RBE) in Modeling C^3I Systems: A quantitative approach," TECOM Test Technology Symposium, Laurel, Maryland, April 1991.

Evolver User's Guide; Evolver Release 2.0, Axcelis, Inc., Seattle, Washington, 1993.

Glover, Fred, Darwin Klingman, and Nancy V. Phillips, *Network Models in Optimization and Their Applications in Practice,* John Wiley and Sons, New York, 1992.

Greenwood, J. A., and M.M. Sandomire, "Sample Size Required for Estimating the Standard Deviation as a Percent of its True Value," *Journal of American Statistical Association,* Vol. 45, p. 258, 1950.

Hald, A., *Statistics with Engineering Applications,* John Wiley and Sons, Inc., New York, 1952.

Kirk, Roger E., *Experimental Design: Procedures for the Behavioral Sciences,* 2nd ed., Wadsworth, Inc., Belmont, California, 1982.

Morris, Scott A, "Electromagnetic Environmental Effects Test Facility (EMETF) Threat Simulator Capabilities," US Army Electronic Proving Ground, Fort Huachuca, Arizona, October 1987.

Natrella, M.G., "Experimental Statistics," Handbook No. 91, National Bureau of Standards, US Department of Commerce, 1966.

Ostle, Bernard, and Richard W. Mencing, *Statistics In Research,* 3rd ed., The Iowa State University Press, Ames, Iowa, 1975.

"Test Officer's Handbook," US Army Electronic Proving Ground, Fort Huachuca, Arizona, November 1990.

Walpole, Ronald E., and Raymond H. Myers, *Probability and Statistics for Engineers and Scientists,* 2nd ed., Macmillan Publishing Company, Inc., New York, 1978.

C. Working Group III - Use of Prior Information in Test Scope and Sizing

Chairman:

Mr. Jim Duff
Technical Director
Commander, Operational Test and Evaluation Force

Co-chairs:

Dr. Marion Bryson, FS
Technical Director
US Army Test and Experimentation Command

Dr. Duane Steffey
Study Director
National Academy of Sciences

Introduction

While the acquisition strategy continues to undergo rigorous scrutiny, it is clearly evident that the future will incur a dramatic decrease in the force structure and result in a reduced requirement for the generation of numbers of new weapon systems. This in turn has the potential to impact T&E in that less dollars will be available for allocation to the T&E infrastructure and process. Accordingly, we must consider new options that enable us to continue to conduct T&E that will provide effective and suitable systems to the war fighter in an economically feasible manner.

The purpose of Working Group III was to address the use of **Prior Information in Test Scope and Sizing** and determine the impact on the broader issue of "How Much Testing Is Enough." The working group addressed four related questions dealing with the use of prior information:

- How should prior information be used to help determine the size or duration of testing?
- Under what conditions can information be pooled or combined?
- Can early DT and late DT information be pooled meaningfully?
- Can DT and OT information be pooled meaningfully?

Working Group III included 32 individuals from the military, civilian government, and private industry sectors. The majority of the members were professionals in the T&E community; although there were clear differences of perception between those who considered themselves "testers/evaluators" and those who were "analysts." The blending of these different cultures resulted in excellent information exchange and frequently drove spirited discussions that benefitted all members of the group. Everyone in the group had the opportunity to express their opinions on the issues being discussed and clear up misconceptions held by others. This free and open discussion helped the group to reach a common definition of terms and provided a meaningful exchange of ideas on each question addressed by the working group.

Approach

The working group chair decided not to break the group into subgroups but rather to keep the group intact and allow all members to contribute to the discussion for each question posed. In retrospect, this turned out to be a wise decision as it forced all members to focus on the issue of **prior information** from a macro view point and attack the issues of *How, When*, and under *What Conditions* prior information can be pooled or combined from a joint tester/ analyst perspective.

Working group sessions were divided into three distinct phases. Phase one focused on the scope and applicability of the four questions assigned to the group. It included three formal and one informal presentations to provide a jumping off point for group discussion. The second phase provided briefings of two case studies where prior information was used or combined and promoted dialogue on how we might build on lessons learned from those cases. The third phase was used to expand discussion on any significant issues noted during earlier discussions and synthesize the results for presentation as part of the group's report.

Topics and case studies presented to the working group are listed below. Abstracts of the formal presentations are provided at the end of this section.

TOPIC AND PRESENTER

"Structured Analysis Approach to OT&E," Ms. Sharon R. Nichols.

"Can DT and OT Information be Pooled Meaningfully? Of Course--Not!," Dr. Carl T. Russell.

"A Bayesian Approach to the Meta-Analysis of Army Field Test Data," Mrs. Kathy Pearson.

"Can DT and OT Results be Combined?," Mr. Phillip E. Wralstad.

"Can Early DT and Late DT be Pooled Meaningfully?," Dr. Alan W. Benton.

INFORMAL PRESENTATION

"Acquisition Streamlining," Mr. John Lyons.

Prior Information

From the ouset it became clear that there were as many different definitions and perspectives of **prior information** as there were working group members. To some, prior information meant information contained in the MNS, COEA, ORD as well as manufactures design information, engineering knowledge, user knowledge, program schedule and cost data, etc. To others, prior information was interpreted as test data that was derived from earlier testing of the same system or a similar system, or from training, exercise, or field data from the same or similar system; or data obtained through modeling and simulation for the system under consideration. Each definition can be correct depending on whether the prior information is being used to help design and plan the test or to supplement/combine with expected test data to reduce testing requirements. The group identified the following broad categories as examples of prior information that would be considered in answering the questions posed to the working group:

- OT/DT/Manufacturers/Foreign test data for the same system.
- OT/DT/Manufacturers/Foreign test data for a similar system.
- Prior test strategy.
- Engineering knowledge/Expert opinion.
- Combat/Exercise/Field Data for same/ similar systems.
- Modeling and simulation.
- Requirements documents.
- Program schedules and cost information.
- The law (congressional/ legal language).

How Should Prior Test Information Be Used to Determine the Size or Duration of Testing?

Dr. Duane Steffey led the group discussion addressing this question. To stimulate the discussion, he provided his perspective on such issues as: (1) What kind of prior information should be used? ; (2) Where do you get prior information? ; How do you use it? ; and, Can statistics help do it better? The presentation achieved the purpose of creating group discussion as many varying opinions were expressed, all based on how individuals interpreted the meaning of the term **"prior information"** (see previous discussion). During these discussions, it became apparent that there were two different schools of thought concerning the issue, depending on whether one's background was as a tester/evaluator or an analyst. Testers/evaluators generally saw prior information as test data from earlier tests of the same system, including modeling and simulation, which could be used to meet test matrix data requirements thereby reducing the amount of testing required. Analysts viewed prior information as those known elements such as criteria from the COEA process, results of component testing, distribution data from similar systems, etc. that could be used with analytical tools to better define the scope and duration of the system under consideration. These differing perspectives quickly forced the discussion "into the weeds" over terminology and led the group to the realization that the terms needed to be clarified before proceeding.

Eventually the group figured out that the issue was not so much the definition of prior information or even the source but rather the bottom line requirement that the *information must contribute* to the knowledge level of the tester/evaluator to improve the T&E process.

Despite the inherent differences between DT and OT, the membership felt the need for these two communities to work closer together, starting earlier in the acquisition process and overcome the "we vs. them" mentality that has long prevailed. It was acknowledged that the two communities needed to better share the information they had before either party could figure out how to apply that information to improve the overall test process.

Several members of the group suggested that they had not had much luck in archiving data and that previous attempts to obtain prior test information had not been successful. Storing the data without standardized protocols for identification and retrieval was viewed as a major barrier to using prior information. One must also know where the data came from (conditions, environment, etc.,) and what assumptions and conditions were involved before one can figure out how to use the data. In all cases it was agreed that the ability to use prior information was dependent on many factors that must be carefully weighed on an individual case by case basis, but that the sharing of existing, relevant data was the necessary first step.

In summary, the group concluded that prior information could be used in the following manner to determine test size or duration:

- To supplement current T&E (DT & OT) when the information is determined to be of sufficient purity that it can be combined with current test data and collectively analyzed and evaluated.

- To complement current T&E results in a comparative or complementary fashion that adds confidence to the analysts determinations when limited or sketchy data was available for the current test. The prior information may have been collected under different conditions or with a somewhat different system configuration, but the results were sufficiently relevant to support use as complementary data.
- In lieu of planned T&E when the prior information is determined to be sufficient to answer the questions posed by the planned T&E.
- To conduct sensitivity assessments to determine the range of test conditions that should be accommodated in the planned testing.
- In accommodating statutory requirements in cases where issues such as environmental concerns must be addressed.
- In an iterative test planning process that must factor in information needs and resource constraints. Particularly in those instances when funding available for T&E is insufficient to fully satisfy tester/evaluator desires and a determination of acceptable risk must be made.
- To focus current T&E through smarter, forehanded planning based on information and lessons learned from prior T&E and system's progress through the acquisition cycle.

Under What Conditions Can Information Be Pooled or Combined?

Working Group Chairman Jim Duff led the group in trying to come up with the answer to this question. To stimulate thought and discussion, Sharon Nichols provided a presentation on a *Structured Analysis Approach* to OT&E. Her presentation introduced an analysis approach that is being selectively tried at AFOTEC. The method uses *Object Oriented Analysis* to develop an "information model" of the operational test concept through the pooling of simulation data. Additionally, Kathy Pearson presented a case study of a series of Army tests where prior information, including expert opinion, was used to determine the expected degradation of unit performance in a chemical environment.

These presentations were successful in causing the group to explore reasons for wanting to pool information and the benefits to be gained. It was agreed that pooling could be used to shorten the acquisition process, save money, reduce test time, and to get state of the art technology into the field faster. There was extensive discussion about the "conditions" under which it would be acceptable to pool or combine information. Most members of the group felt that the issue of pooling or combining information needed to be addressed early on in the T&E planning process to ensure the various tests were designed in such a manner to allow the pooling or combining of data to occur later. Several members expressed the frustration that the question concerning the conditions under which data could be pooled or combined, if ever asked, was typically too late to influence the test process.

Two types of data were discussed as appropriate for pooling or combining: (1) specific test data from a prior phase of testing (DT or OT) of the same system; and (2) related test data from an older, but comparative system. As a result of discussions during this session, it was decided that the pooling or combining of information depended on a number of *conditions* that had to exist either independently or in combination:

- Was polling/combining of data called for in the TEMP? For the most part members felt that it was necessary for the TEMP to address the issue of what information was to be pooled/combined and how it was to be treated.
- When test conditions were similar, members felt it would be appropriate for information to be pooled/combined. Conditions did not have to be identical but alike enough to give the evaluator confidence in the outcome.
- When there had been limited configuration change to the system between test phases, it would be appropriate to consider pooling/combining information from one phase of testing with another.
- When MOEs/MOPs were of like enough definition to be viewed as interchangeable.
- When then is an accredited model, data gathered in the modeling and simulation process can be pooled/combined with actual test information and vice versa.

During this discussion period there was an interesting issue that raised for the second time in working group deliberations. The issue concerned the use of prior information from the COEA process in DT and OT. It was suggested, from the analytical perspective, that the COEA should be used as a derivative source of information for determining test scope and sizing and that the MOEs developed during the COEA should be used for later testing. There was strong disagreement from the testers/evaluators in the group about using COEA information for any purpose other than as a factor to be considered. The concerns expressed were that the COEA information was premature and did not necessarily reflect the final system requirements. Testers/evaluators reiterated that they should only test to the requirements of the ORD and if something in the COEA was important enough it should be carried forward to the ORD. Concern was expressed that the use of the COEA information and MOEs might indeed lead to too much testing beyond that which was necessary to determine the system's operational effectiveness and suitability. As an example, the use of "Battle Outcome MOEs" from the COEA was seen as a situation where test size and scope could be driven far beyond what was considered sufficient to determine the effectiveness and suitability of a specific system.

Can Early DT and Late DT Information Be Pooled Meaningfully?

Can DT and OT Information Be Pooled Meaningfully?

Because of the similarity between these two questions, it was decided to combine them into one discussion issue led by Dr. Marion Bryson. The discussion was kindled by presentations by Dr. Carl Russell, Dr. Alan Benton, and Phillip Wralstad, addressing the pooling of DT and OT information. Additionally, a case study providing an example of the use of prior information and the pooling/combining of information was presented by John Lyons.

Many felt that the issue of sharing information must be addressed before any consideration of pooling or combining could be made. It is broadly perceived that there is an unwillingness between members of the DT and OT community to share their information with one another. This obstacle must first be overcome if there is to be meaningful progress in pooling/combining information in future testing. Many felt that the inherent differences between early DT and late DT, and DT and

OT would make it difficult to pool information. Nonetheless, the majority also felt that we should continually look for opportunities to pool information, and, on a case by case basis, determine if conditions, definitions, configuration, data requirements, test analysis, etc. are similar enough to support pooling. It was clear throughout the discussion that analysts generally felt that they had the necessary tools to support pooling/combining of information in more instances than the testers/evaluators felt comfortable with. Extensive discussions on the use of analytical techniques such as simple aggregation, META analysis, Bayesian statistics, and nonparametric statistics, with several examples of each were provided as potential means for the meaningful pooling of information.

Throughout the discussions, it became clear that the term "pooling information" also meant different things to different people. For ease of understanding, it was suggested that the pooling of information could occur by several methods:

- **Mingle** - Treat the data as if they all came from the same population/distribution.
- **Compare** - Treat similar data side by side to address and explain the magnitude of differences.
- **Allocate** - Identify data to be collected in one test environment to eliminate redundancy.
- **Combine** - Eliminate OT or later DT when there are no specific test issues that haven't been addressed in other testing. Perhaps give an "OT flavor" to some DT events.

In summary, it was determined that information may be able to be pooled meaningfully depending on the specific circumstances of each case. Issues such as the test structure, parameters examined, test environment, system knowledge, and the evaluator's judgment must all be considered when determining if pooling can occur.

The dominant theme throughout the discussion was the need for the testers, evaluators, analysts, DT community, and OT community to get together early in the acquisition process and ***organize*** for the pooling of information and address the How, When, and What issues in the TEMP.

Major Recommendations

Modify DoD 5000.2 to Address The Use of Prior Information.

The use of prior information in test scope and sizing has definite merit. During this mini-symposium, it became apparent that there certainly is justification for considering using prior information in the test process. The extent of this role is dependent on many variables, some of which were touched on in Working Group III. While not applicable to all situations, it is appropriate that provisions for the use of prior data be addressed in governing directives. DoD 5000.2 should be modified to make provisions for the use of prior data and provide policy guidance as to when, under what conditions, and to what extent this use would be acceptable to the decision maker.

Modify DoD 5000.2-m to Allow the TEMP to Address the Use of Prior Information.

The TEMP is the single document that is agreed to by the sponsor, the testers, and the decision authority. As such, the TEMP should

be the document where the use of prior information is spelled out for the specific system under consideration. Likewise, it is the mechanism for coordinating test planning to ensure that information to be pooled/ combined is of sufficient purity to satisfy varying test phase objectives. DoD 5000.2-M should be modified to require the TEMP to address the How, When, Where, and Why issues related to the use of prior information. Approval of the TEMP with this level of detail will ensure the decision makers concurrence with the use of pooled/combined information from other sources.

TEMPs Should Be Developed to Ensure That DT and OT Data Can Be Efficiently and Effectively Combined Where Possible.

On a selective basis there is a place for the use of prior information in determining test sizing and scoping. Those in the best position to determine how best to pool/combine the prior information are those persons charged with preparation of the TEMP. Accordingly, the program manager and the operational tester must work together in TEMP development to ensure that the TEMP approaches the issue from the most effective and efficient manner.

TEMPs Should Identify Which Elements of OT and DT Data Can Be Combined and How.

This recommendation is a follow-on to the previous one and would require that drafters of the TEMP identify specifically which elements of prior information they consider candidates for polling/combining, and how they would propose effecting the action. It would also require a description of the analytical techniques to be used in selective instances and would offer the approval authority a clear understanding of the agreed upon data elements to be pooled/combined and the conditions under which the pooling/combining would be accomplished.

Other Recommendations

The Use of Prior Information Should Be Part of an Iterative Process That Occurs Within The TIWG . . . Early On!

Provisions for using prior information in test scope and sizing must be organized very early on in the acquisition cycle. The determination of how, how much, when, where, etc. must be made through an iterative planning process that begins as early as possible in the cycle, balancing information needs with resource constraints. This process could best be coordinated through the TIWG.

Increase The Focus on The Use of Statistical Methods For Using Prior Information.

While the use of prior information seemed to be a difficult issue to many testers, the statisticians in the group professed to have the analytical tools necessary to make it a meaningful and viable part of the process. If prior information is to be used in test scope and sizing, it is imperative on all involved in the process to be more aware of and conversant with the analytical tools available to make informed decisions.

Identify Existing, And Develop New, Statistical Techniques For The Use of Prior Information.

While the tools to accommodate the use of prior information are currently available, not all involved in the test process are aware of

their existence or utility. The techniques currently available must be identified and understood by those responsible for test planning. Where necessary, new techniques must be developed if existing ones are inadequate for the information being considered.

Statisticians Must Go to The Field.

One of the more significant lessons learned from Working Group III was that the statisticians were not cognizant of what the testers/evaluators were doing and the thought processes involved. Likewise, the testers, for the most part, had little understanding of what the statistician could add to the process and what tools he had available to help in test planning, execution, and evaluation. This recommendation encourages the statistician to become more involved in understanding what the tester/evaluator is doing and to get out into the field to observe what actually happens in testing.

Testers/Evaluators Must Step Across The Threshold of The "Magic Room."

This is a corollary to the previous recommendation. The tester, in general, does not have a good understanding of what the statistician can do to help the test process. He must work more closely with his analytical counterpart and develop a better understanding of what tools are available to him.

Data Outputs From New Systems Should Carry Their Own Data Structure Definitions/ Identification.

One of the frequent problems encountered in going back and reviewing past data is the inability to determine the data structure of the information recorded. Usually this renders the data useless as a consideration for test planning or pooling/combining of the data. Future systems should be required to carry their own data structure definition that describes the format, language, conditions, date, location, etc. of the information recorded. Such a requirement will ensure that persons who later review the information will have the information they need to make a determination as to its relevancy to future test planning and test data collection.

Summary

In trying to address the larger question of "How Much Testing is Enough?," Working Group III was tasked with looking at the issue of using prior information for both test scope and sizing and for pooling/combining with other test data. The entering premise was that T&E, as we knew it in the 80s, has become too expensive and too time consuming in this time of reduced budgets and the need to shorten acquisition cycles. This use of prior information to help reduce this cost and time was the genesis of the questions posed to the working group. Dr. Hamre, in his keynote address, stated that the elaborate testing infrastructure that was built in the 80s no longer can be supported on such a grand scale given the budget environment of the 90s. His recommendation was that T&E (specifically OT&E) be imbedded in the acquisition process and not remain an independent activity outside of the process. The conclusion of the working group, albeit conditional, was that the use of prior information—be it knowledge, experience, education, or test data—was a potentially viable tool for reducing test cost and duration.

Before prior information becomes useful in the T&E process, it is imperative that we overcome the **mistrust** that has developed

over the years between the manufacturer, the developer, the testers/evaluators (DT & OT), and the decision makers through the **sharing** of information and **early involvement** of all parties in the acquisition process and test planning. We must **plan for the use of prior information** early in the process and ensure that the information archived suits later use.

It also became apparent throughout the working group discussions that the test community and the statisticians do not have a good understanding of how the other performs his job nor what benefit can be gained by involving the other community in the process from the beginning.

The mini-symposium provided an excellent forum for the exchange of **information** between members of a wide range of communities. It made us all more aware of the concerns and capabilities that others had to offer to the T&E process, and made us focus on the issue of how to best adapt the T&E process to the realities of the 90s. Although a definitive answer to the question of "How Much Testing is Enough" was not developed, clearly there is a place for the use of prior information in the T&E process. The extent of this use is dependent on a number of conditions all of which must be addressed on an individual case basis.

Working Group 3 provided a basis for discussing the use of prior information in test scope and sizing. Initially, everyone realized that there are many categories of prior information. Some examples are OT, DT, contractor and foreign test data for the same and similar type systems. Prior information for sample size (or duration) testing may be used to supplement current T&E, to complement current T&E, or in lieu of planned T&E. Information can be pooled or combined in programs that have similar test conditions, similar test results, a limited degree of configuration changes, like definitions, and when there is an accredited model. Pooled data may be analyzed as if it all came from the same population/distribution and similar data may be treated side by side to address and explain the magnitude of the differences.

Major recommendations (clearly implementable) include:

- Modify DoD 5000.2 to address the use of prior information.
- Modify DoD 5000.2-M to allow the TEMP to address the use of prior information.
- Develop TEMPS to ensure that DT and OT data can be efficiently and effectively combined.
- Ensure that TEMPS identify which elements of DT and OT data will be combined and how.

Appendix 1. ***A Bayesian Approach to the Meta-analysis of Army Field Test Data*** **- Kathy Pearson Northwestern University CSC Professional Services Group**

Abstract

The U.S. Army has conducted a number of operational tests in the last two decades to determine degradation in unit performance of certain combat tasks under the threat of enemy chemical weapons employment. In particular, the "Combined Arms in a Nuclear/Chemical Environment Force Development Test and Experimentation" (CANE FDTE) program has conducted four tests that measured unit performance in a chemical warfare environment. The overall purpose of the CANE program has

been to "provide measured data and determine how well combat and support units can perform their missions in extended operations where nuclear and chemical weapons are employed" [Independent Evaluation Plan for a Combined Arms in a Nuclear/Chemical Environment Force Development Test and Experimentation (CANE FDTE), Revision 1.5, October 1988]. In response to requests from other members of the Army community for performance degradation data, the U.S. Army Chemical School has now recognized the need to synthesize these results into a single range of degradation values to make the results more useful. These requests have come from a variety of sources, including combat modelers, combat developers, and trainers.

This paper presents the development of a methodology for obtaining a single range of estimates for the expected percent difference in performance of a task in chemical warfare conditions. The methodology incorporates all of the information available on human performance of combat tasks in a chemical environment, including the subjective judgements of military experts. Specifically, a probability distribution is obtained for the percent difference in unit task performance by aggregating both the field test results and the subjective assessments of military experts, as well as any other data from appropriate sources such as actual combat data or field exercise data.

The proposed methodology incorporates principles of meta-analysis and Bayesian statistical techniques to obtain the distribution. First, expert assessments are elicited to determine a prior distribution, representing the "prior knowledge," for the expected percent difference in performance of a particular combat task. Next, the field test results of unit performance of the task are treated as observational data and combined mathematically with the prior distribution to obtain a posterior distribution for the expected percent difference. This posterior distribution represents the synthesis of both subjective and experimental data, and provides the ability to not only give point estimates of the expected percent difference in performance, but also ranges and confidence intervals of the expected difference.

Appendix 2. ***Can DT and OT Results Be Combined?*** **- Phillip E. Wralstad, TEXCOM IEWTD**

Abstract

As fiscal and personnel resources become increasingly constrained, and emphasis is placed on streamlining and reducing the length of the acquisition cycle, one of the possibilities that emerge for examination is the combination of information from different types of tests. The paper addressed particularly two testing sequences—developmental testing and operational testing. Definitions in guidance for each type of testing were reviewed to point up differences and similarities. Instances where data from developmental testing and operational testing could not be combined, and where combination was possible, were cited. The circumstances which made combination acceptable are identified. The paper concluded with observations on factors which, in the author's view, bear on whether responsible combination of data or results from developmental testing and operational testing of a system can be beneficially combined. These include the structure of the issues, the parameters being examined, understanding of the system and its concept of use, execution and conditions or setting of the test, the judgment of the evaluators, the need for exhaustive and precise analysis, the acceptable degree of

uncertainty in regard to answers for essential issues, and the risk the decision makers are willing to accommodate. If combination is to be done, early agreement to the approach by the involved elements of the acquisition community is essential, and needs to be followed by early planning to produce test settings which support subsequent combination; combination after the fact tends to overlook important differences. The question of combination of data from different sources remains one that requires resolution on a case-by-case basis.

Appendix 3. ***Structured Analysis Approach to OT&E*** **- Sharon R. Nichols, HQ AFOTEC/ SAN**

The presentation introduced an analysis approach which is being tried by some Operational Test and Evaluation (OT&E) test programs at AFOTEC. The method uses Object Oriented Analysis (OOA) to capture the "information model" (like a blueprint) of the operational test concept. This is a new application of OOA, which was originally developed as a software design tool. It is useful in operational test applications because it provides a focused view of the problem to be solved.

The purpose of this paper was to act as a catalyst for discussion of the questions, "Under what conditions can information be pooled or combined?" and "When and how can modeling and simulation data be combined with operational test data?"

To start with, we first looked at why we use OOA or some kind of analysis method. When the members of a test support group receive an assignment to develop a test concept/plan, they are generally overwhelmed. Initially, it appears that everything in the world relates to their program and interacts with it. After reviewing the Operational Requirements Document (ORD), the System Threat Analysis Report (STAR), the Mission Need Statement (MNS), etc., they begin to home in on the key "Objects," their attributes (descriptive information about the object which is important to the problem under study), services (transformation behavior of the object), states (different forms of the object throughout its life cycle) and relationships between each other. The relationships break down into Whole-Part (one object "is made up of" another), Generalization-Specialization (a specialized object "is a" general object type), and a general association. A Collection of objects with identical attributes, services and states is called a class.

At the highest level, the Class/Objects are grouped into the following subjects: ENVIRONMENT, SYSTEM UNDER T&E, AND TEST CONCEPT DEVELOPMENT MINI-BRIEFS. In the background, in order to store data about these subjects, we set up a data base schema which reflects Air Force OT&E policy, e.g., Critical Operational Issues (COIs) relate to Measures of Effectiveness (MOEs), which are made up of Measures of Performance (MOPs), etc. The next level break down below the Subject Level Diagram, looked at the categories of information (class/objects) for three of the five MINI-BRIEFS which AFOTEC requires as a part of the test concept development process. One was the REQUIREMENTS mini-brief. The REQUIREMENTS class/object is associated with SYSTEM, MOE, and ISSUE class/objects. Next, the RESOURCES class/object is associated with the TEST SCENARIO, PERSONNEL RESOURCE, MATERIAL RESOURCE, and LIMITATIONS. Lastly, we looked at the mini-brief dealing with the

FOCUS OF TEST class/object. It was here where we found the classes and relationships that I believe are most applicable to look at to answer the question about combining Modeling and Simulation with test data. We need to look closely at the OPERATIONAL MISSION SCENARIO and TEST SCENARIO conditions as well as the LIMITATIONS and CONSIDERATIONS for each MOE.

Appendix 4. ***Can DT and OT Information Be Pooled Meaningfully? Of Course--not!*** **- Carl T. Russell, TEXCOM Experimentation Center, Fort Hunter Liggett, CA**

Abstract

Pooling information from DT and OT is typically suggested as a way to gain statistical significance without increasing test costs. Usually, total miles or hours of operation planned in DT and OT are added together and pushed through the exponential distribution to claim that enough testing is planned. Alternatively, fancier statistical methods of combining data are proposed. These formal approaches to pooling information are unjustified and misleading. An informal approach, however, can be quite helpful. Intuitively, all tests on the same system during the same period of system life provide related information which must be considered together in any cost effective acquisition process.

Appendix 5. ***Can Early DT and Late DT Be Pooled Meaningfully?*** **- Dr. Alan W. Benton, U.S. Army Material Systems Analysis Activity**

An important consideration in reducing test requirements when evaluating system performance is pooling information from various tests and test phases. More specifically, can early DT test results be pooled with late DT results meaningfully?

First we need to answer what we mean by early DT? Late DT? General categories of DT range from research (before phase 0) to technical feasibility (phase 0) to ATD to engineering development (phase 1) to EMD (phase 2) to PPQT (prior to MS III) to first article (procurement) and so on. Most would consider the range to cover at most the last three, perhaps even just the last two.

One evaluative area, reliability, has utilized a unifying management and estimation process called reliability growth, monitoring and measuring progress as configuration changes are made to the system under development. While other evaluative areas lack similar statistical underpinnings, it is believed that the process is applicable. The process consists of design-test-analyze-corrective action-fabricate-retest.

In a study conducted to compare plant and field test results for reliability, a number of factors were found which precluded combining results, which if they had been considered in test planning, could have resulted in pooling of "early and late" testing. Among those reasons for differences were:

- improper stress loading during plant tests - both environmental and software loading
- typically only one environment (cold, hot, etc.) in field versus many in plant
- possible additional environments in field - sand, dust, rain
- automatic test equipment used in plant
- inaccessibility of unit under test in plant
- cables, connectors not exercised in plant

- differences in "system" under test - government furnished equipment may not be included in plant testing

Examples of systems where results could be pooled/not pooled were given. In addition, examples were given to illustrate that tests were not repeated in late DT when successfully passed earlier and significant configuration changes were not made.

In summary, can early DT and late DT be pooled meaningfully? It depends. It depends on similarity in test conditions, configurations, methodology, test results, etc. In order then to be able to pool results and reduce testing, these and other planning factors need to be considered up front early in the evaluation and test planning process.

D. Working Group IV - Practice and Theory

Chairperson:

Mr. Jim Baca
Defense Evaluation Support Activity, (DESA)

Cochairs:

Dr. William Lese
OASD(PA&E) (GPP/LFD)

Dr. Jim Streilein, USA Material Systems Analysis Activity (AMXSY-R)

Mr. Ron Jacob
46th Test Wing, Eglin AFB, FL

Rapporteur:

Mr. William Sieg
QUADELTA INC

Background

This group was tasked to take the "practitioner" approach in investigating "How Much Testing Is Enough?" The basis for the task is the potential to apply structured evaluation methodologies to minimize the amount of testing performed. Essential to these methods is choosing the right design of new experiments, incorporating intermediate decision points in the sequence of tests, and collecting data that can be shared by all interested parties. The tasking to the working group was directed at understanding the practical difficulties in accomplishing such an evaluation.

The focus of the working group centered around three questions posed by the mini-symposium steering group:

- "What are the limitations of design of experiments methodology as applied to individual tests?";
- "When are sequential testing techniques appropriate?"; and
- "How can detailed test definition issues, such as choosing scenario choices (number and type), player uncertainty, number of replications, and baseline testing be addressed?"

Working Group Structure

The working group was divided into three panels. Participants were randomly and equally divided and were rotated through each panel. Thus, each participant was able to address all the topics. All the participants later met in a combined session to review and discuss major issues, observations and recommendations.

The potential of too much overlap occurring among the three panels was discussed

among the cochairs before the conference. The concern was effective use of time by not repeating discussions already covered by another panel. Therefore, Panel A was asked to focus on higher level effects that inhibit use of structured test methodologies; Panel B was asked to investigate the application of specific methodologies; and Panel C was asked to include the use of modeling, simulation and other modern technologies in detailed test definitions.

To stimulate thoughts on structured methodologies, a briefing, "A Simple Decision Aid for Determining Initial Test Size", was presented by Mr. William Moore, USA TEXCOM/ IEWTD.

Participants were encouraged to interact and express their views on the issues. The working group included representation from the top to the bottom of the testing community. Ideas were allowed to emerge and not be stifled for reasons of rank or position. The approach was successful with near 100 percent active participation in the discussions.

Discussions at the three sessions were, as expected, quite different. Nevertheless, common threads emerged. The following is a synthesis of what was discussed.

Structured Evaluation Techniques

Test and evaluation is essentially a learning process. To be effective, Test and Evaluation (T&E) practitioners must continually review what they know and what they do not know about the system under test (SUT).

Many tools exist to assist an evaluator in structuring a test. These include design of experiments, Taguchi, nonparametrics, simple single variable sample sizing, sequential sampling, sequential testing, application of steps of good "statistical" test planning, etc. The collection of these tools can be thought of as a toolbox that can be drawn from depending on the specific application. However, there is no single tool or technique that is considered to be universally applicable. Application of these tools requires a discipline that can only be achieved by management commitment to a well trained work force.

Overall Observations

The working group did not try to reach consensus on the following observations. However, enough individuals expressed similar views to warrant their discussion here.

a. *Knowing how much testing is enough depends on the needs of the customer(s).* Of paramount importance is clearly establishing who the customer(s) is and what he or she wants from the T&E. A T&E organization may have more than one customer and their requirements may vary. For instance, the primary customer may be a decision maker in the acquisition community who is trying to support a production decision, while an operating command customer may be interested in rapidly integrating the system. What satisfies one customer may be too much or too little for another. The right amount of testing is a function of satisfying the critical customers and maximizing the system knowledge for others.

b. *More effective linkage of test planning, data collection and reporting is needed.* Participants felt opportunities exist to reduce testing in many programs through greater cooperation among analysts in the DT, OT and FOT communities. Significant strides

have been made by some services in getting these communities involved at the inception of programs to more clearly state requirements. Some Test Working Groups apparently work better than others in improving cooperation, sharing critical resources and information, and reducing redundancies. Enough participants in this working group however, were still concerned about the loss of corporate knowledge, the in-practice effectiveness of working groups and the potential to more effectively apply coordinated structured test design and evaluation methodologies across the test phases.

c. *T&E must quickly adapt to the changing environment in DoD.* New demands and challenges facing the T&E community must be addressed. The impact of Advanced Concept Technology Demonstrations (ACTDs) and new acquisition policies on T&E is not sufficiently understood. Changes in the acquisition process will probably require quicker response and greater flexibility by the T&E community. Also, the T&E community should take advantage of the new acquisition policies and new technologies to do its job more effectively.

d. *There are few examples of Distributed Simulation (DS) being successfully applied by the T&E community.* DS is being widely discussed in different forums, but there are significant education and technical hurdles that must be overcome before DS becomes an effective T&E tool. Concern was expressed that DS may be incorporated into T&E before the limits of its effectiveness are understood.

e. *Models and simulations (M&S) are not necessarily the answer to the high cost of testing.* For various reasons, M&S is not now the primary source for T&E data. In some cases, M&S could be more expensive than traditional field testing. There may be real benefits in the future, but for now, there are no easy answers.

f. *Field Testing is already a pretty lean business.* There is not really a lot of fat to trim in the number/extent of field tests performed on many systems. Data coming out of the family of tests however needs to be more efficiently and effectively collected and applied to satisfy the needs of the set of customers.

g. *In many cases, T&E is working all right but, the value of T&E is not being adequately publicized.* As evidenced by Desert Storm, testing has been a major contributor to the highly effective combat forces we have fielded. Although we want to find ways to save money, we still must avoid a hollow force. Our national military strategy commits us to the maintenance of highly effective combat forces. The only way we can confidently sustain such a force is by affordable T&E that addresses the needs of both the development and operational communities.

Recommendations

Integrate evaluations

Emphasize integrated "Evaluators" subgroups to CT/ DT/ OT/FOT Test Working Groups to discuss and design the structured methodologies and analytical products in all phases of T&E. The purpose is to eliminate expensive testing by understanding requirements up front and evaluating available information before additional testing is designed. This "Evaluation and Test (E&T)" focus would encourage better linking of data and analytical products/results and should satisfy the set of customers with minimum testing.

Require the design of analytical products with sample data to be briefed to customers before the start of a test

If the customers clearly understand what products are intended to be delivered at the end of the T&E, they could save considerable resources by eliminating products of no use to them. Also, the T&E agency could more accurately design the right evaluation and be more timely and effective in producing its products.

Refine DoD 5000 to include more common terminology and policy/guidance on T&E of systems largely employing ACTDs, COTS, NDI, etc.

Differences in terminology are a catalyst for other more serious differences in a T&E program.

The T&E community should take advantage of advanced analytical tools and instrumentation. T&E leaders should take the initiative to ensure that technological advances are incorporated into T&E resources to give customers what they need

Train/equip/support the T&E work force

Highly qualified and motivated people are crucial for effective T&E. Managers should aggressively strive to avail their people of new tools and professional courses dealing with structured evaluation methodologies.

Look for good examples of DS for T&E

Capture Anti-Armor Advanced Technology Demonstration (A^2ATD) lessons learned, both good and bad, from the Army and use them to give others a starting point to examine DS applications.

Panel A Observations, Issues, and Recommendations

General:

Panel A concentrated on the structure of T&E and other high level effects that determine whether an environment is conducive to the application of structured T&E methodologies. It was within this context that the panel considered its question.

General Observations:

- Testing differs among services.
- Hardware and software differ in time development.
- There are no pat answers on how much testing is enough.
- Large portions of T&E functions well.

Issues:

a. *Better linkage is needed through the entire T&E process.* T&E has been fragmented in many ways. This is true for the *process of doing T&E, the documents* that are generated, the *data* that is collected and even the *organizations* that do the work.

1) *Better linkage is needed in the process of doing T&E.* Better synergies should be sought between DT and OT. This does not necessarily mean a merging of DT and OT, although some advocate that. It does mean that DT, OT and FOT should be approached in the same way. The regulations that govern the conduct of T&E

even institutionalize some of the problems. For example, the format of a T&E Master Plan (TEMP) has separate chapters for DT and OT, encouraging more separation in practice than is required. The results of DT should be the basis of OT tests, differing mostly in scope.

2) *Documents should build on one another such that nothing is or needs to be repeated.* There is a need for common terminology among the participants.

3) *Better sharing of data is needed across all phases of testing.* Evaluations performed at various stages (milestones) should take advantage of all the data that exists at that time. Data base structures should be standardized, at least within each program. Then, from the first laboratory test to the end of the program, a common data base is available.

4) *Better linkages among T&E organizations is desirable.* Developers could benefit from early involvement of both developmental and operational testers. Concepts such as the Army's test integration working group (TIWG) should be considered.

b. *A perception that OT&E is just a final exam needs to be changed to OT&E being a viable part of various decision processes.* Testing, acquisition and operational use of a system are intertwined from the cradle to the grave of each program and cannot be separated. This raised the issue of making operational testers part of the acquisition reporting chain. Some felt this would bias their objectivity. Others felt that OT&E is already held hostage to the budget, some of which is controlled by advocates of the program. Still others felt that operational testers can and are already objectively involved. The common concern was that OT&E should be recognized as being done to learn the operational factors of a system in support of various decision processes and not be merely viewed as a pass or fail test.

c. *Funding*. Program managers have severe funding limitations that often dictate their course of actions. T&E is therefore often faced with tight schedules and seemingly inflexible program managers. With shrinking budgets, program managers and test directors are under more pressure to establish the right balance between development and T&E. Compromises are made which affect the application of structured methodologies.

d. *Program vulnerability*. Shrinking funds also make programs vulnerable to further cuts, or even cancellation. Program managers react by creating a team that supports the program even more strongly. This advocacy discourages testing to learn, because literally any bad news found by testing threatens the program.

e. *New concepts are affecting T&E policy*. T&E policy has not kept pace with new concepts in the acquisition process. These include distributed simulations, advanced technology demonstrators, horizontal technology insertion, and battle laboratories.

f. *Accounting and cost breakout.* The true cost of testing is not clear, principally because it spans multiple program elements.

Industrial funding of testing facilities adds to the difficulty because it has substantial institutional funding that is not apportioned to individual programs. The problem leaves planners in a quandary, not knowing where to streamline. In one example, a panel participant claimed that 90 percent of the dollars expended for testing occurred after the production decision. Others found that figure unbelievable. Whatever the number, the effect of unclear cost breakouts is apparent.

g. *Cost effectiveness of T&E.* Finding and fixing problems before Milestone III has a 10 to 1 cost benefit over fixing problems after Milestone III. However, probably since they involve different organizations (R&D vs User) and different "color" money, the acquisition community may not have sufficient incentives to stress pre-Milestone III T&E.

h. *5000-series publications.* Participants felt that solutions to some of today's problems are already institutionalized in these regulations, but that they are not being followed. Thus compliance is an issue. Others expressed views that the regulations are outdated, do not reflect current thinking in the T&E community and revision is necessary.

Recommendations

The recommendations that follow generally parallel the issues above. While there are no organizations identified for implementation, the level of the issues suggests high levels within the DoD.

Institutionalize T&E linkage throughout the acquisition process

As discussed above, more effective linkage is possible in the T&E processes, documents, organizations and data. Institutionalization will take more than just changing the regulations; it will take the commitment of all testing organizations.

Charter a formal group to identify and develop a cost accounting breakout for T&E

This action would resolve issues regarding where the testing dollars are going and would suggest where efficiencies are possible.

Charter a study to assess the impact of new concepts on T&E policy

The study would address at least the concepts mentioned above under issues.

Focus on the high payoff of pre-Milestone 3 testing

Specifically emphasize evaluation in the concept definition, design and early development phases to both correct deficiencies and reduce the amount and cost of system level testing.

Require programs to more closely follow DoD 5000 publications.

Revamp the 5000-series publications

Bring the regulations more in line with current practice.

Revamp the format of the TEMP

Remove the systemic separation between developmental and operational testing.

Panel B Discussion, Observations and Recommendations

Introduction

The panel first had to agree on the area of concern and the meaning of the terms in the issues the panel was asked to address. The scope was defined as the T&E application of "Statistical/Engineering/ Operations Research Techniques" to provide the "best info" to decision makers during the acquisition cycle of military equipment hardware/software (HW/SW) as a critical step in providing the field user with a required capability. Multiple terms in this scope were also discussed to establish a common terminology.

Definitions

Statistical/Engineering/Operations Research Techniques: There are many techniques in the toolbox of the T&E practitioner. These include design of experiments, Taguchi methods, nonparametrics, simple single variable sample sizing, sequential sampling, sequential testing, application of steps of good "scientific" test planning, etc. However, the best test will only provide a set of data under specified conditions. There needs to be an evaluation to synthesize the collected data into information that can answer the relevant questions. Examples of analytical tools for estimating, transforming and extrapolating data include: analysis of variance (ANOVA), regression, response surfaces, decision analysis, fuzzy logic, analytical engineering models, and modeling and simulation (stochastic, deterministic, hardware-in-the-loop [HITL]) to mention a few.

There may be several steps in translating the raw test data into information usable by the evaluator and/or decision maker. The evaluator may use detailed test data, modeling/ simulation (M/S) data, engineering analysis, expert/panel judgments, etc., to develop specific relevant findings and recommendations which will assist in reaching decisions. To do this, it is critical to know the importance of the information to the decision makers and the range and accuracy of the information. Then an evaluation plan can be developed which includes all the sources of data: test, analysis, M&S, fielded systems, etc.

Required test data must be carefully selected in each test instance. Considerations in test design include the marginal costs and potential value, possible real time review, and analysis/use plan. It is important to keep in mind that data maybe of limited future use due to test specific circumstances or system configuration changes, but these should and must be considered in the design process to minimize testing while making maximum use of test information.

Overall Findings: The panel came to the following conclusions:

- There is no single tool or technique that will be uniformly and universally applicable (see discussion on specific phases).
- The application of specific techniques from the toolbox will not result in information at no cost.
- It is possible that we may not be doing enough T&E now.
- The trend is toward technologies, threats, and instrumentation that are more complex and expensive; without smarter T&E, this could potentially increase, not decrease, T&E costs. The penalties for future wrong decisions may be far more costly than the savings that accrue from reduced T&E. With budget reductions, we may risk getting no system or a bad system.

- Appropriate application of the toolbox through the life cycle of development will result in improved decision "information" and systems for the field user.
- Panel members provided examples where the toolbox is in use to some extent and with some success.

Limitations: The following reasons for limitations of toolbox use were also discussed:

a. There is a lack of resources to do even "minimum" T&E. Resources include number of samples, time, dollars and qualified T&E personnel.

b. Late involvement of T&E personnel limits options.

c. There are "bad" requirements. Perhaps these are untestable or unreasonable requirements which preclude a valid evaluation but which T&E must still address.

d. Some panel members reported instances of apparent lack of decision maker concern for real information; ignorance of or uncertainty in implications of the recommendations.

e. Many panel members indicated that T&E personnel had not done an adequate job of communicating and justifying resource requirements and the real risks if these requirements are not met.

f. There is an apparent lack of trust among T&E organizations which hinders the sharing of data.

Recommendations

The T&E and acquisition communities should develop (and use) a life cycle "continuous integrated" (CI) approach to T&E of systems across all phases by using the toolbox in combining M&S and physical testing where practical.

This CI approach should use appropriate tools at each phase to get the "best information" from available resources to allow the required decision to be made.

Senior leadership should empower the tool users by providing the physical and financial resources to effectively use the tools, by emphasizing tool box use and by encouraging improvement and expansion of the toolbox. There need to be experts who have the training and resources to carry out this task.

For each system, an integrated T&E team should be formed that is responsible for requirements testability, integrated T&E planning using the toolbox, and plan justification to decision makers.

Ensure T&E team members are adequately trained in toolbox use. Current training for T&E and acquisition personnel focuses on general management

The panel felt that toolbox education and application should be a top priority issue. Most people come into T&E completely ignorant of such tools as DOE and thus approach their problem with a "one step at a time" mentality.

Management and technical leadership should become acquainted with the toolbox, particularly DOE. They should require input on the use of the toolbox and consideration of uncertainty and risk.

Discussion of Some Tools

The panel's focus questions were difficult to discuss thoroughly in the given time and panel composition. However, the following discussion of several tools across life cycle phases is offered as an initial presentation for further development. As requested, we focused on design of experiments (DOE) and sequential testing and sampling.

Scientific Test Design - All Phases

No matter what technique from the toolbox is used, there are steps in scientific test design that should be followed:

- Define questions/information needed.
- Determine data parameters needed.
- Develop an analysis/evaluation plan.
- Determine important factors/levels (model-physical, empirical, hypothesized, expert judgment).
- Determine required accuracy.
- Estimate expected variability (conduct pilot test).
- Determine the required sample size.
- Select the factors/levels and order of presentation of trials.
- Conduct the experiments as designed.
- Review data as soon as possible.
- Conduct the analysis/evaluation as planned.
- Document and publish for the T&E community the lessons learned throughout the whole cycle.

Sequential Tools

One tool from the toolbox is "sequential testing." We define sequential testing as making decisions along the test cycle at planned points based on testing to date. This approach is close to current practice. We inferred instead, that we were to address "sequential sampling."

Sequential sampling is defined as stopping or continuing a test by making decisions (pass/fail/continue) during testing under uniform conditions based on observed results. This distinction (between sequential testing and sequential sampling) is relatively small but does have significance. Sequential sampling can be explained as in Figure 1.

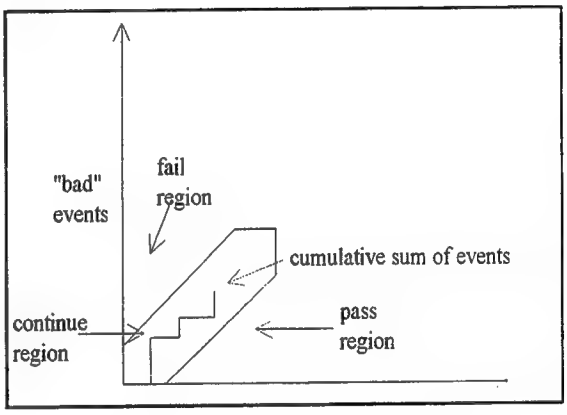

Figure 1. Sequential Sampling

Prior to the start of test, the whole region and the decision rules must be set up following statistical rules and they must be rigorously followed. (One panel member commented that Bayesian techniques exist that allow for changes during the test.)

Design of Experiments

One extremely useful tool is design of experiments (DOE). (Taguchi test designs will be considered as a very sparse test design and limited subset of DOE with some peculiar analysis approaches.)

DOE is a formal approach to selecting factors to be varied, setting the levels of the factors, and determining the levels and the number of repeats of each sample. DOE has evolved into a very powerful and useful tool. Included with DOE are formal approaches to translation of data to "information" by analysis of variance, analysis of covariance, regression, response surfaces, etc. Thus, there is a determined approach to setting up a test and deriving information from the results. These techniques;

- build a mathematical model of a physical situation, allowing one to "estimate" results between "levels" which are random (versus fixed) in nature
- measure compliance with a requirement
- allow interpolation or extrapolation of results
- measure significant differences between products or treatments
- permit selection or improvement of system design to develop a robust product (of which Taguchi and his followers have been strong advocates).

Model-Test-Model

We also discussed the model-test-model (MTM) approach to T&E and its relation to "practice and theory" of test design as a key new tool in combination with other tools. We diagramed a flow process with feedback loops as shown in Figure 2.

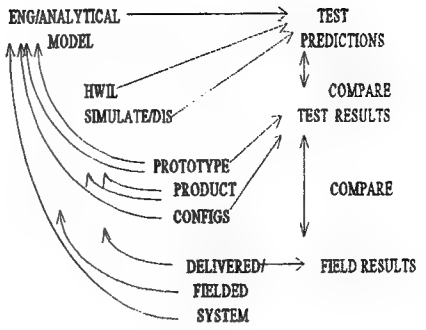

Figure 2. Flow Process with Feedback Loops

The panel participants generally supported the MTM approach as an integral part of good evaluation. However, we also discussed many critical issues with MTM to include: Cost Benefit/Resources Needed, Extent of Use, Fidelity/ Wrong Decisions, VV&A Resources, and Unknown Unknowns.

Integration Modeling and Physical Testing

Panel members generally supported integration of M/S and physical testing. This approach allows evaluation of M/S validity and development of decision support information. In very general terms, there is the expectation that M/S can be used to conduct many more "simulated" tests over all factors and levels of interest. "Physical" tests using the toolbox can then be concentrated at a much smaller number of strategic areas. Such an approach may increase risks due to the uncertainty of models, but may be the most cost effective approach.

Tool Use by Phase of Life Cycle

Concept Exploration

Ideally, DOE should be initiated in Phase 0. Here we should start M/S as part of the MTM. DOE can assist in building and proving the model and to initiate robust product design. The T&E team, with early involvement, should consider testability, reliability, maintainability, produceability, supportability, and requirements evaluation. However, DOE does not have wide enough acceptance in the acquisition community and most likely will not be used effectively. The DOE concept requires that the entire series of tests be designed prior to the start of testing and be applied consistently. But too many people have not been trained in DOE techniques and therefore focus on one event at a time. There is great pressure to "demo" effectiveness; not to gain the insight via rigorous evaluations that could preclude later problems. Limitations on time and available resources also hamper the ability to effectively apply DOE techniques.

Product Development

During product development, DOE should be used on a more advanced or complete level. Benefits include building and improving models for MTM, robust designs to improve products, the measurement of the accomplishment of goals and the potential to reduce unnecessary repetition of tests. The payoffs to using DOE are great. The techniques however, need to be better accepted.

"Sequential" tests may be used to combine results with emphasis on using "precise" variables data perhaps versus qualitative data. Of course, there are limitations and factors to be considered: training, resources/ time/ contract/proprietary data rights prior to government ownership.

Product Verification

In the product verification phase, technical testing combines laboratory and open-air or field testing. These tests are combined tests of specific parameters. Evaluations of these parameters can benefit from DOE and sequential sampling. Panel discussions indicated some T&E professionals are using these tools today. In this phase, DOE is used to build models (MTM), test hypotheses, and improve products. Again there are difficulties/limitations including too many factors, levels and sources of variability. There are known unknowns and unknown unknowns and as always there are limited resources (time, money, samples).

There may be limited control of many factors: weather, terrain, threats, personnel, etc. Sequential sampling is used sometimes. However, the Program Manager (PM) normally wants to stop when failures occur, change the system, and restart, which moves outside the statistics of the plan. It is difficult to convince PMs of the need to plan/program more time/resources/samples, although statistically speaking sequential sampling on average will use less time/samples/resources than fixed length tests of the same magnitude. To effectively use sequential testing, quick data analysis and decisions are necessary.

Production Qualification Testing

This is where whole system testing occurs. Here the evaluation should include all factors based on specific parameters and use M/S. In many instances, results from the previous phases of testing can be used.

Checking must be done to ensure that production of a good design has not degraded previous results and that required corrections have been made. Sub-designs of previous testing should be considered and supplemented with MTM.

Initial Operational Test and Evaluation/Operational Evaluation (IOT&E/ OPEVAL)

Considerable discussion was devoted to the conduct of IOT&E/OPEVAL, the "final exam" in the development program. There are specific legal restrictions and requirements:

- Production representative hardware
- Representative units (operators, etc.)
- No contractor involvement.

The OT community normally requires free play or limited control events where operational factors (some unknown) can still influence results. Few large exercises can be run. Some participants questioned the real need for IOT&E and whether results are different from technical testing with representative soldiers. There is a real need to OT the whole system and for force-on-force testing. The discussions centered on the use of DOE or sequential sampling because of the nature of such testing. Some argued that there are too many unknown and/or uncontrollable factors to really use DOE.

Panel C Issues and Recommendations

General Discussions

Panel C concentrated on the application of modern techniques and technologies to make detailed test designs more effective. Issues discussed included modeling and simulation, both local simulation and distributed simulation, linking the test ranges, and the need to manage and promote insertion of new technology into the T&E toolbox. The following is a summary of the three subgroups' discussions.

Issues

- Testers, in general, have insufficient experience with and knowledge of DS, particularly distributed interactive simulation (DIS) as used by the training community. Insufficient attention has been paid to test methodology suitable for DS. Scenarios, data structures, and the categories suitable for DS need to be developed.
- Modeling is extensively used in T&E, but several areas need improvement. Models constructed in a hierarchial architecture need better linkage. Before the models are used they need to be put through a rigorous T&E process.
- Several issues were raised concerning the Major Range and Test Facility Base (MRTFB). Participants felt ranges are not being sufficiently linked in support of T&E nor, apparently, is enough being done to determine the requirements for doing so and no one appears to be driving the ranges to link. There is at least one initiative to develop a demonstration system. While this issue overlaps the DS issue somewhat, some of the linking envisioned is as simple as allowing a test force at one range (or other location) to view or control a test at another. The ranges were also criticized for not moving faster in inserting new technology, such as GPS, into their instrumentation systems. Often when they do, specific customers are charged for the develop-

ment and the use of the new technology. This often constrains test customers to use what is available, rather than what is needed.

- Advanced Technology Demonstrations (ATD), and their derivatives, are not sufficiently discussed in the 5000 series of OSD documents. Their relation to T&E and the effect on T&E is not well understood. Rigor in the design of ATDs could provide a foundation of data used to evaluate the resulting system thereby reducing the overall cost of T&E.

- Can modern technology be used to reduce the number of replications in testing? While another of the panels approached this from the Design of Experiments viewpoint, this panel approached it from application of models and simulation in test design. For example, if the physics of the penetration of armor by a particular type of warhead is well understood, it may be possible to predict penetration as a function of standoff distance. If a particular warhead of that type is tested and agrees well with theory, only a few replications may be necessary to confirm that the warhead is or is not performing as it should. This led to a discussion of the more general use of modeling and simulation to reduce the amount of more expensive open air range testing. It was noted that there is an increasing tendency on the part of the services to use digital simulation, HITL, and installed systems tests to lead up to open air range testing.

Recommendations

Charter a study of test use of DS.

This needs to be a proactive study that emphasizes how to test using DS. As a minimum this study should recommend categories of tests suitable for DS, methodologies that take advantage of DS, and the network system requirements of T&E as opposed to training.

Develop a strategy or plan to insert new technology in the test process

The OSD CTEIP program is a big step in the right direction, and the efforts under the new T&E Executive Agency structure put together by the services should help.

OSD should investigate current T&E policy and guidance to determine if ATDs can be better accommodated in acquisition policy.

Systems should spend more time in DT&E before moving to OT&E.

There usually needs to be more time early in DT&E for the testers to help the Program Manager mature the system and define the limits of the system performance envelope before more rigorous testing. Systems are usually tested only to the limits and specifications of the ORD and COEA, rather than to the true limits of the system. It was also recommended that DT&E be used more effectively prepare the system for OT.

Other Discussion Topics

The following issues and recommendations represent less widely held opinions but nevertheless contributed to the overall quality of the discussions. They are stated here without discussion as possible areas for future discussion.

The limitations on use of contractor involvement during operational tests reduces opportunities for using existing data.

Independent OT is called for in legislation, not necessarily based on mistrust of the acquisition system; but because of a need to adequately replicate the operational environment. The discussion centered on the application of structured methodologies and the OT benefits of increased use of DT generated information.

The potential for reducing testing costs exists in combining operational testing and training. In the Navy, that is becoming the norm for the later stages of OT. There may be potential for savings from the other services applying this philosophy.

Issues:

- Decline in numbers of technical personnel.
- Lack of experts in software testing.
- Need for generic test facilities.
- Different approaches among services.
- No linkage between COEA developers and OT testers.
- TEMP should play to the A-spec.
- Can test resources contribute to the industrial base?
- Suitability testing not done as well as it should be.
- There's an inconsistency of definitions and terms among the players in the acquisition process.
- "Testing community" needs access to pre-MS-0 data.
- Some TIWGs are debatably submerged in administrative details and not effective as intended: to integrate activity.
- Flexibility is needed in the testing and acquisition systems to handle changing technology.
- Milestone 2 is increasing in importance, 3 is decreasing.
- In addition to acquisition decisions, testing should support operational employment, training and other purposes.

Recommendations:

Regarding Modeling and Simulation, use validated M/S to support testing, integrate M/S into test process, consider more R&D of process modeling.

Form a group to investigate where the testing dollars are going.

Create repositories of data from testing.

Post-production, surveillance (R&M) test data should be fed back to the DT/OT community.

Coordinate analytic products (to be developed by testing) with the decision maker before testing.

Determine clearly what customers (decision makers, users, etc.) want to know -- then work on how much testing is enough.

Chapter IV

Synthesis Group

Chairperson:

Mr. Ed Brady, FS
Strategic Perspectives, Inc.

Co-chairs:

Mr. Clayton Thomas, FS
Chief Scientist, AFSAA/ SAN

Dr. Marion Williams, FS
Chief Scientist, AFOTEC/CN

Dr. Ernest Seglie
Scientific Advisor, DOT&E

Mr. Art Fries
IDA

Mr. Eugene Visco, FS
Director, U.S. Army MISMA

Dr. Patricia Sanders
OSD(PA&E)

Introduction

The goal of the MORS/ITEA Mini-Symposium on "How Much Testing is Enough?" was to identify key issues for and develop insights into the promotion of a more cost effective test and evaluation (T&E) process in support of the acquisition of defense systems. The symposium included a panel discussion (by distinguished present and past members of the T&E and acquisition communities), a keynote speech by the Comptroller for the DoD, presentations in a plenary session, papers and discussions in four separate working groups, final reports from the working groups, and an overview summation from a synthesis group of experienced T&E practitioners. On the last day of the symposium, preliminary conclusions and recommendations were briefed to all attendees and to invited high-level Service T&E officials. These focused on specific ways to assess how well the T&E resources are used, methodologies for improving the design of T&E programs, and approaches for enhancing the timeliness and utility of T&E products.

One of the constant top priorities in the recent T&E experience has been the execution of rigorous tests and the generation of analytically sound evaluations. In the past this quest for quality occasionally led to prolonged and expensive T&E programs. Never was inefficiency explicitly encouraged or condoned, but, nonetheless, its perceived pervasiveness has prompted some to suggest that there ought to be less testing now that resources are more limited. Given current and projected acquisition funding levels, however, it is quite likely that the information derived from T&E will play an even more important role in supporting the decision making process. Fundamentally, there are fewer dollars to go around and the inherent risks involved in making incorrect decisions become increasingly magnified. These opposing considerations, i.e., fewer resources but more important, are central to the prevalent impetus for demanding that T&E become more cost efficient, and they clearly defined the critical challenge confronting the symposium.

In response, the symposium drew upon the wide diversity of its many participants and

fostered extensive discussions and deliberations. These culminated in the identification and articulation of numerous recommendations for improving the cost effectiveness of the evaluation process and the individual tests themselves. The recommendations centered about the utilization within the T&E process of the following:

- overarching cost-benefit analyses,
- integrated test and evaluation teams,
- comprehensive and efficient planning,
- early operational assessments,
- modeling and simulation, and
- efficient and innovative techniques in the design of experiments.

Complete statements of the associated recommendations are presented and elaborated upon in the Executive Summary below, as well as in the subsequent summaries for each working group. The self-ascribed charter of the synthesis group, tasked with writing the Executive Summary, extended beyond merely compiling the major points reported by the individual working groups. First, the working groups interacted synergistically and a coherent summary of their respective concerns and products necessitated some consolidation. In addition, the members of the synthesis group elected to interject their own personal perspectives and emphases as appropriate.

It is noteworthy that some of the symposium's recommendations entail the expansion of the present T&E expertise and discipline into challenging new areas—technology demonstrators, modeling and simulation, distributed simulation, and novel acquisition policies and concepts. To keep pace with the changing acquisition climate, the scope of current T&E activities cannot remain stagnant. Instead, it is essential that it be accordingly broadened to embrace and encompass these new areas, and to support the efficient integration of all viable T&E modes into comprehensive programs. Not all of the symposium's recommendations will be easy to implement. Positive progress in these directions is required, however, before significant improvements to the cost effectiveness of defense T&E can be attained.

Summary

The MORS/ITEA Mini- Symposium on "How Much Testing is Enough?" was structured to address several fundamental T&E issues:

- Cost/benefit consideration of T&E programs,
- Optimization of T&E programs,
- Use of prior information in test scope and sizing, and
- Practice and theory of test design.

Each prescribed issue defined the primary focus of one of four distinct working groups, used to partition the symposium attendees according to their expertise and interest. After considerable discussion and debate, the working groups arrived at numerous (often overlapping) conclusions and recommendations. This Executive Summary is a consolidation of the working groups' products, modulated by the perspectives of the individual members of the synthesis group.

A large number of proposed courses of action (i.e., specific recommendations) are embedded within the discussions presented below. They have been collectively grouped under three overall symposium conclusions and three broad recommendations associated with the second conclusion. These can be summarized as follows:

- The T&E process must adapt to the evolving acquisition climate.
- Significant improvements to the T&E process are possible.
  - Operations research principles should be applied to the planning process.
  - Integrated T&E teams should be empowered to facilitate comprehensive and efficient programs.
  - The design, execution, and analysis of individual T&E activities should better exploit existing statistical approaches and techniques.
- The positive momentum established by the mini- symposium must be sustained.

Conclusion 1: ***The Defense System Acquisition Climate Is Evolving - Defense T&E must Adapt Accordingly.***

The environment in which defense systems are acquired is changing dramatically in two key respects. First, current and projected funding levels reflect great reductions. Second, new classes of system acquisition strategies have been introduced recently and will continue to be emphasized. These include, for instance, advanced technology demonstration (ATD), advanced concept technology demonstration (ACTD), horizontal technology insertion (HTI), product improvement program (PIP), and integration of commercial off-the-shelf equipment.

The reduction in funding has prompted the widespread acknowledgment that the T&E process must become more cost efficient. Although fewer dollars and resources may be made available for T&E, it is likely that the demand for quality information and insights that only the T&E process can provide will not diminish. Indeed, the importance of T&E in supporting decision makers becomes even more critical as the risks involved in misallocating sparse funds grow increasingly larger.

The DoD 5000 Series Acquisition Directives long have served as the guidelines under which defense systems are developed, tested, evaluated, and procured. They do not, however, presently accommodate many aspects of the novel acquisition strategies listed above. The scope of the Directives must be appropriately expanded, and the roles of T&E within that framework must be clearly elucidated.

There are thus two basic ways in which defense T&E must change. Foremost, without sacrifice of breadth or quality, it needs to become more cost efficient. Further, it must achieve more flexibility, fully embracing the expediency goals inherent in the wave of new acquisition strategies.

Conclusion 2: ***Steps Can Be Taken to Significantly Improve Defense T&E***

Recommendation 1: ***The T&E Planning Process Should Be Subjected to Fundamental Operations Research (OR) Scrutiny, Including Explicit Consideration of Alternative T&E Strategies, Contingency Planning, and Cost / Benefit Tradeoffs.***

There are many aspects to this recommendation. The foremost of these perhaps is the recognition that at the conception of the initial T&E planning process the basic principles of OR can and ideally should be applied. Certainly some segments of industry and business routinely avail themselves of such analytical tools.

No longer bound by a prescribed rigid and omnibus approach to the acquisition of defense systems, program managers and planners now have the freedom to ponder various alternative development and T&E strategies. Moreover, the recent emphasis on cost efficiency effectively mandates that variants of standard or traditional T&E approaches be examined and detailed cost comparison analyses be undertaken. Alternative T&E concepts should be formulated and evaluated early enough to be inputs into the overall acquisition strategy of the system at the time the budget is being finalized. Otherwise, funding and resource assignments will be made somewhat arbitrarily. The rational specification and study of comprehensive yet viable alternatives can only be undertaken as part of an integrated team effort involving, at a minimum, contractor, developmental, T&E, and user personnel. To mention but one example, the potential for embedded instrumentation to support testing and training should be evaluated as part of the system design and costing.

Currently, a Test and Evaluation Master Plan (TEMP) typically outlines a single sequence of scheduled test events that is "success oriented." There is little or no programmed slack time, and no accompanying discussion or consideration of what is to occur should any major system or test planning problems arise (or, for that matter, if system development and T&E successes proceed much more smoothly than projected). Yet in practice such discoveries, albeit mostly difficulties, often occur. In this sense, the TEMP does not serve as an efficient "master" plan.

A more realistic and useful TEMP would openly acknowledge the likelihood of confronting specific types of problems endemic to the class of systems under development. Additionally, it would build in the flexibility required to incorporate, as need be, contingency plans for overcoming or ameliorating encountered obstacles. For example,. suppose that dedicated contractor testing of a missile system in (laboratory or field) countermeasures conditions reveals system hardware or software deficiencies. It certainly would be advantageous at that time if the previously specified firing matrix (already published in the approved TEMP) could be suitably augmented or otherwise modified to replicate the offending conditions and reexamine missile performance under the troublesome conditions, once system upgrades have been incorporated. Potentially even more cost efficient might be a blend of actual missile firings and extensive modeling studies by contractor and government teams. Without adequate contingency planning, however, neither approach might be economically or expeditiously feasible. By explicitly considering contingencies and factoring in perceived levels of technical risk associated with various testing phases, the cost and value of alternative T&E strategies can be better assessed.

A critical challenge in the application of OR principles to the T&E planning process is the quantification of T&E program costs and benefits. The focus should not be limited to a pre-deployment perspective; a broader life-cycle view provides a more accurate picture of the full extent of costs and benefits. For instance, several mini-symposium attendees reported that large portions of the total T&E costs over the life-cycle of a system often are incurred after the Milestone III full-production decision (estimates varied between 30 and 90 percent). If subsequent investigations substantiate this assertion, information on likely post-deployment T&E costs and benefits should be incorporated into the planning process for pre-

Milestone III T&E activities, as well as into the full-rate production decision making process. The costs of T&E programs often are not well known (e.g., there is no identifiable T&E line in budgets), and the current accounting practices are nonuniform and tend to obscure rather than illuminate actual costs. Moreover, there is no central repository for T&E cost information. Some combination of the T&E and Program Analysis and Evaluation communities should develop a viable work-breakdown structure and associated accounting scheme, with the objective of formulating a rational methodology and comprehensive historical data bases (e.g., consistent measures of projected and actual costs/ benefits) supporting assessments of anticipated T&E costs for any new defense system.

The benefits of testing pose a similar difficulty in quantification. While testers and evaluators more often than not are the victims of poor budgeting, they themselves are typically guilty of not being able to readily articulate the benefits of T&E, either expected from a planned program or derived from an executed program. The OR discipline suggests that plausible alternative T&E strategies should be nominated, with each being costed and critically examined for likely benefit. Potential measures of benefit can be quantitative or qualitative. Program Executive Officers (PEOs) often express risk in terms of bottom-line dollars, but some risk reduction activities are not amenable to such translation —e.g., lives saved, troop morale, and public acceptance are not easily quantifiable. But, in many instances, somewhat crude measures might suffice. In "life or limb" questions, however, we should opt to be very risk averse. It may be quite reasonable to explicitly relate benefit to the time in the system acquisition cycle in which critical information (such as discovery of major failure mechanisms) can become available. For instance, the benefit weighting function might be that finding a failure mode in the design drawings prior to any production is worth 100 units, uncovering it after the commencement of production is worth 10 units, and not having it surface until the system is fielded is worth only 1 unit. Such an analysis could further incorporate the cost of the tests and evaluations to be utilized at each stage.

In summary, from system conception through post-deployment activities, the T&E planning process can take advantage of basic operations research principles that promote greater cost efficiency. In particular, the true costs and the derived benefits of T&E need to be measured consistently and more widely understood. First steps are possible and must be taken. "We keep the improvement up by just ... measuring. If it doesn't get measured, it doesn't get improved." - General Michael Loh, Commander Air Combat Command (from Creating Government That Works Better & Costs Less, Report of the National Performance Review, p. 54, by Vice President Al Gore).

Recommendation 2: ***Each Individual T&E Program Should Empower a Small, Stable, "Integrated T&E Team" (With Some Mix of Contractor, Developer, Trainer, Operational Test and Evaluation (OT&E), User, and Office of the Secretary of Defense (OSD) Representation) to - Manage Design, Evaluation, and Implementation Issues; Continually Monitor T&E Planning Activities and Review Emerging Results; and Revise T&E Plans as Warranted. The Team's Dual Emphasis Should Be on Comprehensiveness and Efficiency.***

The formation of integrated teams is a common theme among management gurus today, and, in fact, major defense acquisition programs routinely establish T&E working groups [e.g., Test Planning Working Groups (TPWGs), or Test Integration Working Groups (TIWGs)]. These have tended, however, to be large bodies comprised of representatives from the Program Manager Office (PMO) and numerous autonomous Service agencies, with each separately responsible for specific action items and particular portions of the TEMP. Moreover, the working groups frequently are not constituted early enough or are too narrowly focused to formulate comprehensive and flexible T&E strategies that efficiently support program milestones. Finally, OSD involvement often begins late in the planning process and its role typically is limited to observer, vice active and coequal participant, status.

The recommendation endorsed here is to build upon the current working group concept - to expand its scope to more directly confront today's T&E challenges, while simultaneously incorporating some degree of standardization and streamlining. Each acquisition program should establish a Test and Evaluation Planning Oversight Group (TEPOG), comprised of representatives from each of the contractor, developer, trainer, OT&E, user, and OSD communities, to oversee and manage the T&E planning process. For major systems, it is recognized that, consistent with the current practice, larger-sized planning groups may be required to attend to the many details of implementation. Under these circumstances, a small TEPOG "executive board", constituted as indicated above, should be established. The key features of the TEPOG would include: early formation, small size, empowered membership, and permanence. The principal responsibilities of the TEPOG would be to: ensure relevant and testable system performance requirements; plan a coherent sequence of laboratory and field tests, modeling and simulation (M&S) activities, and other analyses; design individual tests and evaluations; constantly monitor and review emerging results; and rescope and redesign T&E activities as warranted. Since the ultimate acquisition objective is to place into the hands of the user a system that is both operationally effective and operationally suitable, all T&E activities, even early in the developmental cycle, should strive to incorporate the maximum degree of operational realism and rigor practicable. From this perspective, it is also reasonable to insist that the TEPOG should be chaired by someone from the user or OT&E communities, and the Director of Operational Test and Evaluation (DOT&E) in OSD be represented. The tenure of the TEPOG should continue throughout the completion of all major operational testing, including the Initial Operational Test and Evaluation (IOT&E) and any Follow-On Operational Test and Evaluation (FOT&E) or significant post-Milestone III activities.

Early formation of the TEPOG is essential because long lead times are required for reserving testing facilities, procuring test hardware and personnel, developing and validating required weapon system surrogates and test range instrumentation, and developing and accrediting models and simulations. A small-sized TEPOG facilitates expedience and responsiveness, as does composition by senior level leadership who are empowered to authorize directly, or with minimal external review, specific agreements and courses of action. Ideally, the individual members comprising the TEPOG would retain their positions throughout the entire life of the T&E program (or as

long as their military assignments permit), fostering both stability and corporate memory. To this end, TEPOG membership should include both civilian and military personnel, and detailed documentation of TEPOG decisions with accompanying rationale should be encouraged. Note that it is quite likely, indeed prefererable, for individual members to serve on multiple TEPOGs. It is important for the TEPOG to retain its leadership and oversight roles beyond the completion of IOT&E, since the Services typically conduct a considerable amount of post-IOT&E testing. Estimates of the total T&E costs over the life-cycle of a system that are incurred after the Milestone III full- production decision have been reported to be in the 30 to 90 percent range. In addition to the Service-designated FOT&Es, the Services also execute numerous types of related testing, e.g., Production Verification Testing (PVT), Enhanced Producibility Production (EPP) Testing, Qualification Test and Evaluation (QOT&E), Qualification Operational Test and Evaluation (QOT&E), Pre- Planned Program Improvements (P3I) Testing.

One of the initial responsibilities of the TEPOG should be to review the Cost and Operational Effectiveness Analysis (COEA). Large acquisition programs, i.e., acquisition category (ACAT) I programs, are required to prepare a formal COEA, to illuminate the relative costs and benefits of alternative system approaches, in support of individual milestone decision reviews. The TEPOG should review the COEA from two perspectives—first, to ensure that the measures of effectiveness (MOEs) used by the COEA appropriately characterize the operational effectiveness and suitability of the system; and, second, to ascertain whether the MOEs are directly testable. For those MOEs that are deemed not be directly testable, the COEA should identify appropriate testable measures of performance (MOPs) and establish how changes in these MOPs relate to changes in the MOEs. The TEPOG should continue to monitor the status of the COEA updates, particularly as the developmental system matures, T&E results emerge, or other program changes (e.g., a revised threat) become evident.

The TEPOG also should play a similar review role for the system Operational Requirements Document (ORD), prepared and approved within the Service and subsequently validated either by the Joint Requirements Oversight Council (JROC) or by the initiating Service. Testability, relation to clearly expressible MOEs and MOPs, and relevancy in light of the most recent program developments are all fundamental issues. Furthermore, the ORD should define the intended operational environment and the minimum acceptable operational performance required in sufficient detail and scope to provide a meaningful basis for future evaluations. The timing of the initial TEPOG review should support the ORD validation process.

The primary function of the TEPOG should be to define a coherent T&E program of laboratory and field tests, M&S studies, and other analyses, that comprehensively address all of the developmental and operational issues while efficiently utilizing resources, including information. The first step of this process entails laying out a chronological set of T&E activities for insertion into the TEMP, with the accompanying schedule of required testing resources and dedicated personnel to be delivered and/or reserved. The initial structure of the activities should be sufficient to verify adequate hardware and software developmental progress, provide sufficient flexibility to accommodate any potentially necessitated

T&E program changes, and furnish early insight into operational performance capabilities. As discussed further below, linkage between the various T&E activities should be explicitly established and noted in the TEMP—including criteria for progressing to later tests, results expected to support the detailed planning of future tests, and opportunities for sharing data and information across different activities. Although results from developmental testing generally are viewed as potential inputs for the planning and evaluation of operational testing, feedback from OT&E to DT&E is also possible. For instance, operational testing may quantify the relative degree to which particular engagement conditions occur and influence the design of the Live Fire Test and Evaluation (LFT&E). Likewise, operational testing may identify particular system performance shortcomings that can be more completely addressed by follow-on detailed technical testing (e.g., countermeasures susceptibility).

The TEPOG should continually review emerging T&E results, to determine whether the T&E program should adhere to existing plans or, as expeditiously and prudently as possible, revise them accordingly. This is of particular importance if system performance shortcomings become evident and/or as the formal OT&E phase approaches. The TEPOG constitutes an ideal forum for communicating, early on and at the appropriate Service and OSD levels, the demonstration of likely system difficulties, their potential ramifications, and possible alternative recourses.

Each individual T&E activity should be designed to provide information for influencing subsequent action, with the set of possible actions being prescribed in advance. Examples include proceeding to the next planned T&E activity according to schedule, and expanding, reducing, or consolidating the scope of future T&E activities to account for specific results or uncertainties demonstrated to date. The actions associated with particular T&E outcomes must be identified in advance to objectively determine the type, quantity, and quality of data and information to support the associated decision. Presently, TEMPs, and even Detailed Test Plans, rarely identify the alternative actions being contemplated. Yet, a clear sense of how the choice of potential T&E program options relates to the possible results from an individual T&E activity is a critical design consideration (for that activity, as well as for the entire program). Insisting on the rigor of identifying what will be done if T&E results turn out one way or another will offset the tendency to "test only to comply with some law." Such "compliance" often degenerates into a malicious waste of time and resources that does not provide information useful to any decision. At present, the TEMP is not an integration document; tests and evaluation activities typically have little relation to each other. This is due, in part, because the various chapters and sections are written by separate organizations.

The early conceptual stages of planning for any particular test or evaluation activity typically benefit from a detailed evaluation "crosswalk," beginning with intended issues to be addressed and culminating in requisite supporting data items to be collected. Once the detailed implementation steps of the activity start to take some firm shape, however, likely artificialities, limitations, and other shortcomings generally become acknowledged. At that point, the TEPOG should repeat the evaluation crosswalk process, but this time in reverse, i.e., beginning with the expected available data items, the context in

which they will be obtained, and associated uncertainties (both statistical and otherwise); and tracing back to the issues (actual vice desired) that they support. This type of scrutiny should highlight those individual system performance issues that are not directly testable, in whole or in part, or for which great uncertainty will remain even after completion of the activity. If necessary, complementary analyses or M&S studies can be utilized to demonstrate the possible range of effects introduced by the test limitations. Decision makers should be notified of any major disconnects or concerns, well before the for record test activity is initiated, so that, as warranted, the planned activity can be canceled, delayed, or modified; or supplementary future testing or evaluation can be scheduled. This is particularly important for operational testing, which tends to have higher visibility than developmental testing.

Nowhere is the inadequacy of the current T&E practice, with its lack of linkage across testing activities, more evident than when a system passes development testing and then conspicuously fails operational testing. Examples exist in which the two types of tests did not even consider the same MOEs. Failure to assess the operational implications of developmental test conditions and results is the root cause of such "surprises." Stated simply, development testing and evaluation (DT&E) is not complete without an operational evaluation of the developmental testing. Moreover, as practicable, the maximum feasible extent of operational realism and tactically significant considerations should be interjected into DT&E. For example, individual tests and M&S efforts in the DT&E phase typically have been designed to address narrowly defined issues (such as demonstrating contractor compliance with a prescribed technical requirement); have involved contractor or other nontactical personnel (vice operational military users); have been comprised of simplistic examinations of individual components, subsystems, or systems in isolation (vice investigations of performance of the complete system and associated interfaces); and have been conducted in benign or otherwise nonstressful environments (vice more demanding conditions compatible with actual combat circumstances). None of these characterizations are inconsistent with the traditional intentions of DT&E. Nonetheless, the current quest for efficiency challenges the T&E community, whenever practicable and without sacrificing the original test objectives, to strive to expand the horizons of DT&E and to glean early insights into operational performance capabilities. Such an integration of T&E objectives, across developmental and operational perspectives, can lead to more efficient use of resources, without any degrading the independent evaluation responsibilities of the various Service and OSD T&E agencies and offices. A systematic framework is required to catalog and interrelate the diverse, scattered sets of DT&E results - not only to better describe the current developmental status of the system, but also to establish and refine the scope of the planned subsequent DT&E and OT&E.

A comprehensive understanding of the information content of DT&E, both qualitative and quantitative, also can aid immensely in the finalization of the design for OT&E—for example, in ensuring that any perceived operational shortcomings or weaknesses observed in DT&E are thoroughly examined and evaluated in OT&E; or, in structuring operational test scenarios and conditions to provide system improvements and enhancements, relative to the existing baseline system, the opportunity to be demonstrated. Moreover, an appreciation

of the likely inherent variabilities of system and force performance capabilities under different test conditions, the potential for confounding effects that introduce bias and extraneous variability, and the influence of test peculiar artificialities and limitations is essential in the rational specification of particular conditions, scenarios, missions, battles, vignettes, etc. to be tested, as well as in the determination of specific sample size and test duration requirements to be imposed. Results from DT&E, previous testing of similar systems, and dedicated M&S examinations of possible OT&E variants can provide useful inputs.

Finally, the TEPOG should play the central role in formulating and managing an integrated M&S effort across the extent of the planned T&E program. The M&S associated with the program's engineering evaluations and operations analyses should be designed to form a continuum of model development without duplication or the need for reinvention. The interplay between M&S activities and testing events should be established early on and closely monitored. All contractor models and simulations should be delivered, with adequate supporting documentation, to the government.

The utility of M&S during DT&E is widely accepted. While most acknowledge the potential of M&S to support OT&E, a wide gap exists between envisioned capabilities and the current state-of-the-art. Distributed interactive simulation, for example, has become a common topic of much discussion among defense T&E communities, but there are few examples to date of how it has been applied to OT&E in a cost-effective manner. There appears to be general agreement that the fidelity of existing distributed interactive simulation capabilities limits their current use as a replacement for field testing.

There also appears to be a general consensus that such simulations are likely to be more useful as a test planning tool or for test force training than as a replacement for field testing. At present, one of the greatest challenges appears to be that of faithfully presenting a visual depiction of a simulated scene—including the fine variations of contrast, shading, background clutter, texture, etc—such that its appearance is "real" as seen by the appropriate sensor, frequently the eyes of an operator.

Whether they be in support of DT&E or OT&E, M&S activities are not necessarily either inexpensive or trivial to develop, implement, and maintain. In fact, in some instances, using M&S can be more expensive than traditional testing. Existing models and simulations, originally designed for possibly different purposes or perhaps emphasizing different aspects of system performance, need to be accredited as being appropriate for the intended purpose at hand. The development, validation, and documentation of new models and simulations likewise can be extremely people- and time- intensive. Similar concerns apply to the maintenance and configuration management of large models and simulations. The widespread perception that M&S results are inherently less plausible than field testing outcomes also hampers their potential utility.

Despite the many obstacles that presently confront the utilization of M&S within some aspects of the defense T&E process, the potential benefits remain significant and merit continued attention. Each T&E program should strive to rationally incorporate M&S-based approaches as appropriate—to examine

technical performance issues and operational issues that could not be otherwise addressed; to aid in the design and execution of tests and evaluations; to reinforce, illuminate, and clarify test results; to conduct sensitivity studies, and to extrapolate observed results and provide predictions of future performance (followed up by formal validations whenever possible). The TEPOG should take the lead in ensuring that the application of M&S throughout the program is fully chronicled—including specific models and simulations utilized, summary of results obtained, consequences and contributions of these results, comparisons of results to field test outcomes and other sources of information, and lessons learned. In this manner, particularly when viewed across the diversity of different T&E programs, compelling evidence of the current and potential utility of M&S can be clearly documented.

Recommendation 3: ***More Emphasis Should Be Placed on Ensuring That Each Individual Test and Evaluation Activity Is Efficiently Designed and Analyzed. In Particular, Experimental Design Techniques and Other Established Statistical Approaches Should Be Better Exploited.***

There are many statistical approaches and techniques, both standard and sophisticated, that can be utilized to increase the efficiency and economy of information collection, processing, and evaluation. The primary potential contribution of pursuing the discipline of statistics is to help assure that meaningful and efficient tests and evaluations are planned and conducted, by complementing the individual steps of the overall planning process (outlined above in the discussion of Recommendation 2), rigorously prescribing particular combinations of conditions and circumstances to be examined and the order in which they are to be tested, and ensuring that adequate (but not excessive) sample sizes and test durations are specified. The second type of major contribution that statistics can provide is in the analysis of results, particularly the extraction of sound estimates of system performance and associated uncertainties from messy data.

The statistical field of design of experiments (DOE) encompasses a broad array of well- established formalized procedures for determining the details of the investigative process—which aggregations of variables and conditions to examine, their particular settings and values (e.g., held constant or free to vary), the corresponding numbers of replicates, and the chronological order of test conduct. The goals of DOE are to provide credible and sufficiently precise characterizations of system performance, both for the situations examined directly and possibly, by rational inference, to a larger applicable universe. Designs can be constructed to protect against the effects of potential confounding factors (such as time trends, player learning, etc.), or for accommodating possible departures from nominal assumptions that are too often cavalierly invoked (e.g., standard postulates for underlying statistical distributions, equal variance observations, etc.). Simple aspects of experimental design principles, such as "randomization" (to guard against unknown influential variables) and "blocking" (to reduce the impact of extraneous variability introduced by known influential variables) should be routinely incorporated into the designs of individual defense T&E activities.

Similarly, the most fundamental embodiment of "blocking," namely comparative baseline testing (involving side-by-side or otherwise similar testing under nearly identical

conditions of both the system under development and the current system it is intended to replace) should be utilized whenever practicable, particularly at the more advanced T&E stages such as OT&E. At a minimum, during the latter and most critical stages of T&E baseline comparisons should be undertaken analytically when they cannot be supported directly by head-to-head testing. Testing both the developmental and current baseline system concurrently (which need not necessarily entail equal emphasis and identical numbers of replicates for both tested systems) has several profound advantages. Foremost, it facilitates relative and unbiased comparisons in the absence of absolute or rigorously defensible performance criteria. In other words, it can be directly established whether, and if so to what degree, the new system is an improvement over the old system (not only with regard to types of tests that were conducted, but also, to broader scope evaluations that utilize the test results as inputs). Testing of the baseline system also helps characterize the difficulty and adequacy of the accomplished testing, and places test limitations into a more readily comprehensible perspective. The tangible advantages provided by baseline testing of a "control" have long been universally appreciated across a broad spectrum of diverse T&E communities, including innumerable industrial, agricultural, pharmaceutical, medical, and other scientific research fields. The common counter argument that baseline testing adds to the overall test time and cost generally pales relative to the practical significance of the insights it provides.

In addition to the fundamental principles of the DOE, there exists an extensive inventory of sophisticated and specialized test designs whose objectives are to quantify the effects of a large number of variables of interest while minimizing the total number of test observations. The built-in symmetry of these designs usually requires that sets of variables be tested simultaneously or according to particular prespecified patterns. As such, these families of designs often cannot be easily applied to complicated testing situations (e.g., large-scale OT&Es) that are constrained by numerous inherent limitations and pragmatic implementation concerns. Their systematic employment may be best suited for some types of developmental tests and large comprehensive simulation studies, e.g., COEAs, sensitivity studies, and analytical extrapolations to regimes that cannot be otherwise examined (for, say, cost or safety reasons). This is especially true when the simulations are expensive, cumbersome, time-consuming, or people-intensive.

A statistical viewpoint is essential for the derivation and justification of prescribed sample size or test duration requirements for a scheduled T&E activity. There are two standard statistical approaches for calculating such requirements—a fixed-sample size formulation and a sequential testing framework. Both approaches are amenable in theory, to varying degrees, to the prudent and reasonable incorporation of data or information from previous testing of the same system (under identical or different conditions), testing of similar systems, expectations based on M&S, expert opinion, etc. In common defense T&E practice, however, the integration of results from different sources generally does not factor into the calculation of test sizes.

The traditional fixed-sample size methodologies determine the extent of requisite testing prior to the initiation of for-record activities and commit test management and control to the completion of all of the assigned

testing (and no more) regardless of the observed outcomes (typically to validate standard statistical confidence reporting procedures). They can be considered wasteful when the test results early on appear to convincingly support a particular conclusion, but the remainder of the scheduled testing is obligated to be completed. Moreover, unless specific steps (of the types outlined below) are taken, the common modes of application of these methodologies disregard any other potential sources of useful information that may been previously collected and rely solely on the new data to be collected in the future testing. Consequently, it can become "overly difficult" for systems to formally demonstrate compliance with a prescribed performance requirement at some reasonable level of statistical confidence.

Often the specified test duration becomes "excessively large" in perception, or, to have a reasonable opportunity of passing the test, the true system performance must be substantially better than what it is required to be. The sequential testing framework, on the other hand, does not fix the point of test termination in advance; the concept is to stop testing as soon as statistical conclusions about competing hypotheses for describing the true system performance are adequately corroborated by the data collected to date and attendant analyses. This can conceivably occur either prior to or after the prespecified end of test that would be designated by the corresponding (i.e., with equal statistical chances of arriving at incorrect conclusions) fixed-sample design, but the expected likelihood is for an earlier stopping time. However, greater efficiency in discriminating between alternative hypotheses is accompanied by less statistical precision (since less test data will have been collected) and possible statistical bias in the nominally reported point estimate of system performance. In addition, the classical application of sequential methodologies also is confronted with difficulties in integrating existing information into the test planning process.

While sequential testing, sampling, and analysis are well developed and have been utilized extensively in the medical and biostatistical arenas, there are very few examples of their implementation in defense T&E. Complicated testing circumstances, such as force-on-force trials in OT&Es, are not all conducted under identical conditions and the proper application of sequential methodologies is not straightforward. Furthermore, often the detailed analysis and scoring of test results take considerable time, and there is little point in delaying ongoing test activities in the meantime. Similarly, there is a strong disincentive to stop testing when many personnel and assets have been assembled, at enormous expense and difficulty, in one location for a carefully planned period of time. Under these circumstances, however, sequential methodologies potentially can still provide great cost savings—not by stopping testing early; but rather, once the overall conclusion is foregone, by curtailing extensive and costly data collection efforts to focus on a minimal set of essential data (i.e., just adequate for "testing to learn," vice comprehensively supporting extensive in-depth diagnostic evaluations). Other natural domains for applying sequential methods to defense T&E are repetitive developmental testing under controlled conditions and major program events that involve the planned destructive testing of very expensive systems, e.g., live warhead firings against actual target vehicles, or flight tests of strategic missiles. In these latter situations, the expected gain (i.e., non-loss) of even one item per test can be extremely important, and the total savings over an extended set of tests can be quite impressive.

Economy in minimizing the magnitude of required test sample sizes and test durations can be achieved by properly incorporating data and information from other sources into the supporting calculations. Great care must be taken, however, to ensure the reasonableness of such procedures. In particular, the conditions under which the historical data or relevant information were obtained and the scoring rules utilized therein must be well understood and precisely related to the specific activities scheduled for the upcoming test or evaluation. This would generally depend on the degree to which sufficiently detailed data collection and description programs were imposed and rigorously maintained throughout the histories of related T&E programs. For example, aside from the fundamental question of what constitutes a "failure," numbers of aircraft flight hours between recorded failures of captive-carried air- to-air missiles ordinarily is not, by itself, an adequate summary of the missile's historical on-aircraft reliability. Indeed, true missile performance can vary dramatically with the type of host aircraft, the aircraft station location on which it is carried, the frequency of carrier landings, etc. Moreover, the quality and thoroughness of the different types of diagnostic tools for detecting failures, e.g., built-in-test equipment or other more comprehensive external devices, as well as the relative frequency of their application can profoundly influence the perceived missile reliability.

The "pooling" of available data and information for purposes of reducing the statistical demand for future data can be pursued in several ways. The most direct approach isolates particular aspects of the system performance characteristics under study, matches them up with collections of existing comparable data, and factors the amount and content of the old data into the calculation of requirements for new data. As warranted, old data may need to be adjusted to account for known differences in usage or operating conditions (e.g., by documented K-factors). A second more sophisticated approach supporting test size calculations relies on Bayesian methodologies—a formal mathematical paradigm that additionally accommodates the possible incorporation of diverse modes of relevant information, such as subjective opinion, as well as perceived statistical uncertainties inherent in all of "pooled" information. Bayesian techniques provide great flexibility and potential for effectively minimizing resource commitments, but only when the underlying assumptions are reasonably consistent with reality. For example, applications of Bayesian theories that are consistent with typical textbook examples often implicitly treat each historical data point as equivalent to a new data point, i.e., they do not directly consider the differences in the test conditions and circumstances under which the two types of data are collected. Thus, additional sophistication beyond standard textbook presentations is required.

The intended application of any "pooling" procedures (for test planning or for subsequent analyses of collected data) should always be scrutinized closely, including the undertaking of sensitivity studies to explore the extent to which ultimate conclusions are responsive to the new data, vice dominated by the old data. Simulated data sets, based in part on historical evidence or wholly on artificially generated (but nonetheless plausible) data, provide a straightforward means of producing large numbers of possible test program outcomes and exploring this issue in detail. Moreover, this approach can be exploited in general to examine the likelihood, parameter-

ized in terms of the true unknown performance, that any proposed T&E design (either with or without "pooling") will arrive at specific types of conclusions, i.e., for assessing the expected adequacy of the design. Alternative analysis methodologies and procedures also can be assessed similarly.

In summary, the formal "pooling" of information, for either test planning or analysis purposes, is a difficult and subtle art. The statistical subject matter can be sophisticated, there is little expertise resident in the defense acquisition community, and great care should be taken when pursuing the application of the available methodologies. Moreover, the conceptual appropriateness of "pooling" requires that the data sets in question be truly comparable. Thus it is imperative that test data and related information (e.g., physical conditions, scoring rules, inherent instrumentation accuracies) be carefully archived.

The DOE techniques and statistical approaches discussed above are accompanied by well- established analytical procedures for quantitatively examining the resultant test or evaluation data. In addition, there are numerous other standard statistical tools that can be utilized to obtain valid estimates of system performance and related statistical uncertainties, potentially even when the data are murky and compromised by confounding effects or other test limitations. These include, in part, regression methodologies (e.g., linear, nonlinear, and logistic) for relating system performance to various explanatory variables, robust estimation procedures (e.g., nonparametrics and randomization-based significance levels) for reducing dependence on particular statistical assumptions, and generalized confidence interval procedures (e.g., bootstrap and large-sample approximations) for situations when no exact theoretical result is available. No individual analysis tool is universally applicable or preferable in all circumstances for which it may be appropriate. For any specific analysis problem, a reasonably prudent course of action is to conduct several alternative types of analyses and determine whether the overall conclusion is sensitive to the choice of a particular mode. If so, the relationship between the resultant conclusion and the underlying statistical assumptions inherent in the alternative analysis procedures should be examined, and the plausibility of the assumptions in question should be assessed.

The great majority of the statistical techniques and approaches described above have not been regularly used in the T&E of defense systems, primarily because of a lack of expertise within the community. However, the need for improved statistical efficiency, in both planning and analysis, will become ever more critical as (1) T&E programs are trimmed, and (2) the additional capabilities provided by future generations of systems become relatively more marginal and increasingly more difficult to demonstrate. Thus, there is a compelling need to upgrade and maintain the level of statistical interest, skills, sophistication, and appreciation in defense T&E.

Towards this end, the hiring of professional statisticians and/or the training of designated in-house individuals within all relevant T&E organizations is essential and should be solidly supported by management (including the commitment of funds and resources). Training ideally would go beyond the mere acquisition of statistical knowledge; it should also encompass the development of consulting and communication skills for effectively interacting with the diversity of potential clients and users, and for plainly disseminating statis-

tical results and conclusions to managers and decision makers. Moreover, the growth and maturation of these cadres of statistical excellence should be constantly fostered, by encouraging continual education to keep abreast of new directions and advances, promoting technical interchanges between organizations to share experiences (both positive and negative), and publicizing significant contributions throughout the T&E community. Some specific steps that can be taken in this regard include: (1) regular sponsorship of technical workshops and symposia that focus, at the detailed level, on the application of statistical test design principles and analysis approaches to the T&E of defense systems; (2) creation of a dedicated journal (or reservation of appropriate sections in existing publications) devoted to similar topics; (3) documentation and widespread dissemination of design and analysis "chronicles" that outline for each acquisition program the test planning and data analysis methodologies pursued, difficulties encountered, lessons learned, and benefits derived; and (4) establishment of a T&E methodological "council" to serve as a depository for T&E chronicles, to provide continual input and contributions to relevant workshops, symposia, and journals, and to provide a rapid response technical review capability available upon request.

Finally, the promotion of sound statistical techniques and practices in defense T&E programs cannot be undertaken in isolation; closer links to the professional statistical community will be required. The ongoing study of the National Research Council Committee on National Statistics, Panel on Statistical Methods for Testing and Evaluating Defense Systems, provides an excellent opportunity to initiate such a relationship. Some formal liaison between this panel and appropriate organizational entities of the defense T&E establishment (e.g., the T&E methodological council described in the preceding paragraph) should be implemented.

Conclusion 3: ***the Mini- Symposium Was a Useful First Step in Improving Defense T&E.***

The mini-symposium was successful at two distinct levels. First, at the individual participant level attendees interacted with a wide variety of associates and, no doubt, developed a broader understanding and appreciation of the T&E community's diverse concerns and perspectives. They also personally contributed to the formulation of specific recommendations for improving the T&E process, and, consequently, in the future are more likely to incorporate the spirit of these proposals into their daily activities whenever opportunities present themselves. The positive reinforcement provided by the collective expression of issues and endorsement of recommendations can only encourage and facilitate the implementation of such actions in practice.

At the policy-making level, the symposium's conclusions and recommendations were briefed to the senior echelon of Service T&E officials—those empowered to codify many of the suggestions directly or to seek formal guidance and policy directive changes from the DoD. Moreover, the written record of the proceedings and findings, as documented in this Final Report, enables the symposium's message to be disseminated throughout the upper tiers of the acquisition and T&E communities.

The mini-symposium was a productive first step towards improving the current T&E

process in support of the acquisition of defense systems. A continual progression of more and larger steps (from each of the two directions addressed above) is required, however, to harness and amplify the precious momentum established to date. Each individual in the T&E family can contribute by daily striving to apply the fundamental principles that underlie the symposium's recommendations. Policy makers likewise are responsible for regularly and openly accommodating and promoting such activities. Furthermore, as appropriate, they should pursue the official implementation of procedures and directives that support the symposium's recommendations.

Appendix A

Announcement and Call for Papers

meaningfully? Can DT and OT information be pooled meaningfully?

4. *Practice and Theory of Test Design*
Chair: Dr. John Wiles, DESA, 703-931-2052
Co-chairs: Ron Jacobs; Dr. Bill Lese, ODPA&E, 703-695-0881; Dr. Jim Streilein, USAMSAA, 410-278-6580.

Focus: What are the limitations of design of experiments methodology as applied to individual tests? When are sequential testing techniques appropriate? How can detailed test definition issues, such as choosing scenarios (number and type), ensuring player uncertainty, number of replications, and baseline testing, be addressed?

Synthesis Group
Chair: Ed Brady, FS, SPI, 703-250-6338
Members: Art Fries, IDA; Dr. Ernest Seglie, DOT&E; Dr. Stuart Starr, MITRE; Clayton Thomas, FS, AFSAA; CAPT Coenraad Van der Schroeff, OP-912; Gene Visco, FS, USA MISMA; Dr. Marion Williams, FS, AFOTEC.

Focus: This group will review and synthesize the mini-symposium findings and identify over-arching issues and recommendations.

PRODUCTS

A report containing all formal presentations and summaries of the working group findings, conclusions and recommendations will be prepared and submitted to the MORS Publications Committee for review within four months of the mini-symposium. An Executive Summary briefing will be offered to the MORS Proponents within one month of the mini-symposium. In addition, all significant results of the meeting will be presented to a General Session of the 62nd MORS Symposium and summarized in articles for the MORS *PHALANX* and the ITEA Journal.

PROPONENTS

The mini-symposium will be jointly sponsored by the International Test and Evaluation Association (ITEA) and the Military Operations Research Society (MORS). MORS proponent sponsors are:

- Deputy Under Secretary of the Army (Operations Research) (DUSA(OR))
- Director, Modeling, Simulation, and Analysis, Deputy Chief of Staff Plans and Operations; Headquarters USAF
- Director, Assessments Division, Office of Chief of Naval Operations (N81)
- Director, Program Analysis and Evaluation, Office Secretary of Defense.

FEES

The registration fee for the mini-symposium will be **$150** for Federal Government personnel and **$300** for non-government personnel.

There will be a luncheon with speaker on Tuesday. The cost is **$20.00**. Box lunches will be available on Wednesday to allow working groups to continue meeting. Cost is **$10.00**. Please include payment with your registration fee.

HOTEL

A block of rooms has been reserved at the Williamsburg Hilton, 50 Kingsmill Road, Williamsburg, Virginia. The rate is **$60** for single/double. Reservations may be made by calling 804-220-2500. The cut-off date is **4 February.**

TIMELINES

- ☛ Abstracts for consideration by working group chairs to MORS office **14 January 1994**.
- ☛ Applications to the MORS office **15 February 1994**. (Application forms are attached to this Announcement and Call for Papers.)

CAVEATS

The Military Operations Research Society does not make nor advocate official policy. Matters discussed or statements made during the symposium are the sole responsibility of participants involved.

All attendees and participants are expected to submit requisite attendance forms and to pay the normal registration fees unless specifically waived by the MORS President. There is no waiver or discount for short-period attendance or participation

Acceptance of an invitation to present a formal paper at the mini-symposium implies an obligation by the speaker to attend the symposium, to provide a paper copy of the paper, if requested by the chair, and to submit a timely written disclosure authorization.

The Society retains all rights regarding final decision on the content of the Mini-Symposium Report.

Col Gregory S. Parnell
President

APPROVED:

Pierce J. Johnson
Contracting Officer's Technical Representative

ORGANIZING COMMITTEE

General Chair:
Dick Helmuth
SAIC
8201 Greensboro Drive, #470
McLean, VA 22102
703-847-5587

Technical Chair:
John Gehrig
Director, Army TEMA
Pentagon, Room 3C567
Washington, DC 20310
703-695-8995

Deputy Chair:
Don Greenlee
SAIC
5107 Leesburg Pike, #2200
Falls Church, VA 22041
703-824-5909

Other Members:
Ed Brady, FS
Strategic Perspective, Inc.
7704 Lakeloft Court
Fairfax Station, VA 22039
703-250-6338

Dr. Dave Brown
Army TEMA
Pentagon, Room 3C567
Washington, DC 20310
703-695-8995

Dr. Hank Dubin
USA OPTEC
4501 Ford Ave
Alexandria, VA 22302
703-756-2367

Jim Duff
COMOPTEVFOR
7970 Diven Street
Norfolk, VA 23505
804-444-5197

Dr. Bob Hinkle
ODUSA(OR)
Pentagon, Room 2D278
Washington, DC 20310
703-693-9467

Gary Holloway
USA TECOM
AMSTE-TD
APG, MD 21005
410-278-1315

Dr. Bill Lese
ODPA&E
Pentagon, Room 2B256
Washington, DC 20310
703-695-0881

Dr. Ernest Seglie
DOT&E
Pentagon, Room 3E318
Washington, DC 20301
703-697-7247

Jim Streilein
AMSAA
AMXSY-R
APG, MD 21005
410-278-6580

Clayton Thomas, FS
AFSAA/SAN
Pentagon, Room 1E386
Washington, DC 20330
703-697-4300

Gene Visco, FS
US Army MISMA
Crystal Square 2, Suite 808
1725 Jefferson Davis Hwy
Arlington, VA 22202
703-607-3420

Dr. Marion Williams, FS
AFOTEC/CN
Kirtland AFB
Albuquerque, NM 87117
505-846-0607

MORS/ITEA Mini-Symposium
How Much Testing Is Enough?

TENTATIVE AGENDA

Monday, 28 February 1994

1600-1800	Meeting of Working Group Chairs, Co-chairs and Synthesis Group Members
1700-2100	Early Registration
1900-2100	Panel Discussion John Gehrig, Chair; VADM William Bowes, NASC, LTG William Forster, Military Deputy, OASA (RDA); MG Ken Israel, PEO/C3; Jim O'Bryon, DDT&E, George Williams, PEO/Tac

Tuesday, 1 March 1994

0700-0800	Registration (with coffee, pastries)
0800-0815	Welcomes, Introduction of Keynoter
0815-0845	Keynote address, Dr. John Hamre, USD(C) (invited)
0845-0915	Report on Survey of User's Needs, Jim Duff, COMOPTEVFOR
0915-0945	Report on National Research Council Study, Dr. John Rolph, RAND
0945-1015	Break
1015-1045	Report on Previous Work, TBD
1045-1145	Report on On-going Research, Dr. Don Gaver, Naval Postgraduate School
1145-1200	Introduction to Working Group Charters, John Gehrig
1200-1330	Luncheon with Speaker: MG Ronald Hite, OASA (RDA)
1330-1700	Working Groups
1700-1830	No-Host Mixer

Wednesday, 2 March 1994

0800-0900	General Session: Progress Reports from Working Groups
0900-1700	Working Groups

Thursday, 3 March 1994

0800-0900	Working Groups: Review Final Presentation
0900-1030	General Session -- Report of Working Groups 1-3
1030-1045	Break
1045-1115	Report of Working Group 4
1115-1145	Report of the Synthesis Group
1145-1200	Summary and Closing
1200-1700	Luncheon Meeting of Working Group Chairs, Co-chairs and Synthesis Group

Appendix B

ITEA/MORS List of Attendees

International Test and Evaluation Association
and
Military Operations Research Society
Mini-Symposium: *How Much Testing is Enough?*
Attendee List (03/03/94)

Natalie S Addison
Military Operations Research Society
101 S Whiting Street
Suite 202
Alexandria VA 22304-3483
OFF TEL: (703)-751-7290
FAX: (703)-751-8171

Duane K Allen
Naval Ordnance Center PAC Div
Code 063
800 Seal Beach Blvd
Seal Beach CA 90740-5000
OFF TEL: (310)-594-7544 DSN: 873-7544
FAX: (310)-594-7200

COL Thomas L. Allen
Air Force Studies and Analyses Agency
AFSAA/CC, 1570 Air Force Pentagon
Washington DC 20330-1570
OFF TEL: (703)-695-9046
FAX: (703)-697-3441
E-mail: ALLEN@AFSAA.HQ.AF.MIL

Norman A Anderson
Vitro Corporation
1801 M Sara Drive
Chesapeake VA 23320-1275
OFF TEL: (804)-366-6214
FAX: (804)-420-1275

Joseph J Angotti
NSWC, Crane Division
Code 0547, Bldg 3173
300 Highway 361
Crane IN 47522-5001
OFF TEL: (812)-854-1511 DSN: 482-1511
FAX: (812)-854-3385
E-mail: jja496@ibm-03.nwscc.sea06.navy.mil

Thomas H Antoniuk
Marine Corps Operational T&E Acty
Quantico VA 22134
OFF TEL: (703)-640-3141 DSN: 278-3142
FAX: (703)-640-2472

Capt Harvey Augustine
OSD JTAMS JTF
1431 McGuire Street
Lackland AFB TX 78236-5532
OFF TEL: (210)-671-1905 DSN: 473-1905
FAX: (210)-671-2459

George M. Axiotis
Naval Sea Systems Command
Code SEA-91T
2531 Jefferson Davis Hwy
Arlington VA 22242-5160
OFF TEL: (703)-602-8557 DSN: 332-8557
FAX: (703)-602-0881

James Baca
Defense Evaluation Support Acty (DESA)
2251 Wyoming Blvd SE
Kirtland AFB NM 87117-5609
OFF TEL: (505)-262-4570
FAX: (505)-262-4504

William E Baker
US Army Research Lab
Attn: AMSRL-SL-BV
Aberdeen Proving Ground MD 21005
OFF TEL: (410)-278-6658

LT Earl E Barnes
Operational Test & Eval Force
7970 Diven Street
Norfolk VA 23505
OFF TEL: (804)-444-2949 DSN: 564-2949
FAX: (804)-444-1200

Brian Barr
Institute for Defense Analyses
OED
1801 N Beauregard ST
Alexandria VA 22311
OFF TEL: (703)-845-6928 DSN: 289-1890
FAX: (703)-845-6977
E-mail: bbarr@ida.org

LTC Michael L Bell
US Army Test & Evaluation Mgt Agency
Attn: DACS-TE
The Pentagon, Room 3C567
Washington DC 20310-0200
OFF TEL: (703)-695-8995 DSN: 225-8995
FAX: (703)-695-9127

Robert S Bell
PRC Inc.
468 Viking Drive
Virginia Beach VA 23452
OFF TEL: (804)-498-5692
FAX: (804)-498-5670
E-mail: VA_beach_2_pouch@PRC.com

DR Alan W Benton
US Army Materiel Systems Analysis Acty
Attn: AMXSY-RE
Aberdeen Proving Ground MD 21005-5071
OFF TEL: (301)-278-6625 DSN: 298-6625
FAX: (410)-278-2043
E-mail: abenton@amsaa-cleo.brl.mil

John A Bentrup
CNA
4401 Ford Ave
POB 16268
Alexandria VA 22302-0268
OFF TEL: (703)-824-2579 DSN: 289-2638
FAX: (703)-824-2949
E-mail: BENTRUPJ@CNA.ORG

John M Berry
Ball Corporation
Systems Engineer Div
2901 Juan Tabo, NE
Albuquerque NM 87112
OFF TEL: (505)-262-4636
FAX: (505)-262-4649

Jamno Bhagat
US Army TACO M, Bradley PM Office
Van Dyke Road
Warren MI 48397
OFF TEL: (313)-574-6633 DSN: 786-6633
FAX: (313)-574-7807

Edward C Brady FS
Strategic Perspectives, Inc.
7704 Lakeloft Court
Fairfax Station VA 22039
OFF TEL: (703)-250-6338
FAX: (703)-250-3637

Gene E Brennan
PRC, Inc
Suite 202
111 Howard Blvd
Mount Arlington NJ 07856
OFF TEL: (201)-770-1331
FAX: (201)-770-1283

DR C. David Brown
Army TEMA
DACS-TE
200 Army Pentagon
Washington DC 20310-0200
OFF TEL: (410)-278-4489 DSN: 298-4489
FAX: (410)-278-4116
E-mail: cbrown@apg-9.apg.army.mil

DR Marion R Bryson FS
HQ US Army TEXCOM
ATTN: CSTE-TSA (DR Bryson)
Fort Hood TX 76544-5065
OFF TEL: (817)-288-1057 DSN: 738-1057
FAX: (817)-288-1253

DR Edward S Cavin
Center for Naval Analyses
PO Box 16268
4401 Ford Avenue
Alexandria VA 22302-0268
OFF TEL: (703)-824-2951 DSN: 289-2638
FAX: (703)-824-2949

DR Charles C Chapin
HQ Army Materiel Command
ATTN: AMCRM-CE
5001 Eisenhower Ave
Alexandria VA 22333-0001
OFF TEL: (703)-274-8322 DSN: 284-8322

Jackie K Clark
MCOTEA
3035 Barnett Ave
Quantico VA 22134-5014
OFF TEL: (703)-640-3141

PROF John D Claxton
Defense Systems Management College
Attn: FD-TE
9820 Belvoir Road
Fort Belvoir VA 22060-5565
OFF TEL: (703)-805-2887 DSN: 655-2887
FAX: (703)-805-3183

Capt John P Connolly
DET 4 AFOTEC
4146 E Bijou
Colorado Springs CO 80909-6899
OFF TEL: (719)-554-4105 DSN: 692-4105
FAX: (719)-554-4001
E-mail: CONNOLLYJP@HQ.AFOTEC.AF.MIL

LTC Brent A Crabtree
HQ Department of the Army
ODSCPER, ATTN: DAPE-MBI-A
The Pentagon
Washington DC 20310-0300
OFF TEL: (703)-697-0367 DSN: 227-0367
FAX: (703)-697-7748

COL W. Beaman Cummings
Marine Corps Operational T&E Acty
3035 Barnett Avenue
Quantico VA 22134-5014
OFF TEL: (703)-640-3144 DSN: 278-3144
FAX: (703)-640-2472

Dianne M Cutshaw
MARCORSYSCOM
Code PSA-O
Quantico VA 22134
OFF TEL: (703)-640-4455 DSN: 278-4455
FAX: (703)-640-2168

Debra O Davis
US Army Project Mgr for Instrumentation, Targets, and Threat Simulations
12350 Research Pkwy
Orlando FL 32826
OFF TEL: (407)-381-8775 DSN: 960-8775
FAX: (407)-381-8838
E-mail: davisd@orlando-emh7.army.mil

Roger E Detrick
Naval Air Warfare Center Aircraft Div
FW05
Patuxent River MD 20670-5304
OFF TEL: (301)-826-4277 DSN: 326-4277
FAX: (301)-737-3859

CPT Jerry D Dickerson
Army Logistics Management College
ALMC
Fort Lee VA 23801
OFF TEL: (804)-765-4250 DSN: 539-4250
FAX: (804)-765-4648

William F Diehl
Walcoff & Assoc., Inc.
Suite 400
635 Slaters Lane
Alexandria VA 22314
OFF TEL: (703)-684-5588
FAX: (703)-548-2881

Robert D Dighton
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria VA 22311-1772
OFF TEL: (703)-845-6992
FAX: (703)-845-6911
E-mail: rdighton@ida.org

Lou Dinicolantonio
Reserve Component Automation System
8510 Cinderbed Road
Newington VA 22122-8510
OFF TEL: (703)-339-1814
FAX: (703)-339-1799

Robert O Dizon
Computer Sciences Corporation
ATCCS Sys Engr & Integration Program
1301 Virginia Drive
Fort Washington PA 19034
OFF TEL: (215)-542-5463
FAX: (215)-643-2929

LTC Bruce J Donlin
OASA(RDA)
Attn: SARD-SM
Pentagon, Room 3B465
Washington DC 20310-0103
OFF TEL: (703)-697-8643 DSN: 227-8643
FAX: (703)-697-3827

DR Henry C Dubin
US Army Operational Test & Eval Comd
Park Center IV
4501 Ford Avenue
Alexandria VA 22302-1458
OFF TEL: (703)-756-2367 DSN: 289-2367
FAX: (703)-756-0779

James B Duff
Operation Test & Evaluation Force
Technical Director, Code 00T
7970 Diven Street
Norfolk VA 23505-1498
OFF TEL: (804)-444-5197 DSN: 564-5197
FAX: (804)-445-9174

David W Duma
Operational Test & Evaluation
(Conventional Systems)
1700 Defense Pentagon Rm 1C730
Washington DC 20301-1700
OFF TEL: (703)-697-3891 DSN: 227-3891
FAX: (703)-614-3992

CPT Gregory J Dyekman
USAIS
Dismounted Battlespace Battle Lab
Attn: ATSH-WCB-O
Fort Benning GA 31905-5400
OFF TEL: (706)-545-3165 DSN: 835-3165
FAX: (706)-545-3841
E-mail: DYEKMANG@BENNING-EMH1.ARMY.MIL

William W Dyess
46 Test Wing/OGM
Eglin AFB FL 32542
OFF TEL: (904)-882-5475

William E Eagar
Georgia Tech Research Institute
Suite 1910
1700 N. Moore Street
Arlington VA 22204
OFF TEL: (703)-528-0883
FAX: (703)-528-8419
E-mail: edeagar@gtri.gatech.edu

Robert A Eberhard
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria VA 22311
OFF TEL: (703)-845-6939
FAX: (703)-845-6911

DR James N Elele
USAEPG
STEEP-MT-E
Fort Huachuca AZ 85613-7110
OFF TEL: (602)-538-4957 DSN: 879-4957
FAX: (602)-538-4973

Jim Engel
Defense Test and Evaluation Professional
Institute
521 9th Street
Point Mugu CA 93042
OFF TEL: (805)-989-7947 DSN: 351-7947
FAX: (805)-989-7952

Lawrence A Eusanio
IDA
1801 N. Beauregard Street
Alexandria VA 22311
OFF TEL: (703)-845-6922
FAX: (703)-845-6977

LCDR John H Fenter
COMOPTEVFOR
7970 Diven St
Norfolk VA 23505-1498
OFF TEL: (804)-444-2954 DSN: 564-2954
FAX: (804)-565-8516
E-mail: fenter@tecnet1.jcte.jcs.mil

James P Finfera
Combat Systems Test Acty
Live Fire Vulnerability Dir
STECS-LI-A
Aberdeen Proving Ground MD 21005-5059
OFF TEL: (410)-278-6879 DSN: 298-6879
FAX: (410)-278-6944
E-mail: STECS-LI@APG-EMH5.APG.ARMY.MIL

CPT Timothy Flanagan
Manprint Perscom
TAPC-PLM
200 Stovall Street
Alexandria VA 22332-1345
OFF TEL: (703)-325-4077 DSN: 221-4077
FAX: (703)-325-7927

LtGen William Forster
ASA(RDA)
Attn: SARD-ZB
103 Army Pentagon
Washington DC 20310-0103
OFF TEL: (703)-697-0397

Christine A Fossett
US GAO
Office of Policy, Room 6800
441 G Street, NW
Washington DC 20548
OFF TEL: (202)-512-2956
FAX: (202)-512-4844

MAJ Essex Fowlks V
Test and Evaluation Management Agency
DACS-TE, Room 3C567
200 Army Pentagon
Washington DC 20310-0200
OFF TEL: (703)-695-8995 DSN: 225-8995
FAX: (703)-695-9127

DR Arthur Fries
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria VA 22311
OFF TEL: (703)-845-2364 DSN: 289-1825
FAX: (703)-845-6911
E-mail: afries@ida.org

LT David F Fry
COMOPTEVFOR
7970 Diven Street
Norfolk VA 23505-1498
OFF TEL: (804)-444-2954 DSN: 564-2954
FAX: (804)-444-1200

PROF Donald P Gaver
Naval Postgraduate School
Dept of OR
Monterey CA 93943
OFF TEL: (408)-656-2605 DSN: 878-2605
FAX: (408)-656-2595

John F. Gehrig
HQ Department of the Army
ATTN: DACS-TE
200 Army Pentagon, Room 3C567
Washington DC 20310-0200
OFF TEL: (703)-695-8995 DSN: 225-8995
FAX: (703)-695-9127

Donald L Giadrosich
USAFAWC/OA
Chief Scientist
203 W. D Ave, Suite 400
Eglin AFB FL 32542-6867
OFF TEL: (904)-882-4543 DSN: 872-4543
FAX: (904)-882-2909

John Gilligan
AFPEO/CB
1090 Air Force Pentagon
Washington DC 20330-1090
OFF TEL: (703)-693-8071

William L Goodall
Raytheon Company
Missile Systems Div
50 Apple Hill Drive
Tewksbury MA 01876-0901
OFF TEL: (508)-858-4914
FAX: (508)-858-1502

Hurrol W Goodwin
OSD/DOT&E
Room 3E333
1700 Defense Pentagon
Washington DC 20301-1700
OFF TEL: (703)-695-1565 DSN: 225-1565
FAX: (703)-614-8891
E-mail: Goodwin_Hurrol@mail-host.dote.osd.m

Donald R Greenlee
SAIC
One Skyline Tower, Suite 2200
5107 Leesburg Pike
Falls Church VA 22041
OFF TEL: (703)-824-5909
FAX: (703)-824-5838

Peter P Haglich
Daniel H. Wagner Assoc., Inc.
2 Eaton Street, #500
Hampton VA 23669
OFF TEL: (804)-727-7700
FAX: (804)-722-0249

DR John Hamre
ASD(C)
DoD Comptroller, Room 3E822
1100 Defense Pentagon
Washington DC 20301-1100
OFF TEL: (703)-695-3237
FAX: (703)-614-2378

CDR Christopher L Hanson
COMOPTEVFOR
Code 35
7970 Diven Street
Norfolk VA 23505-1498
OFF TEL: (804)-444-2954 DSN: 564-2954
FAX: (804)-445-8516
E-mail: oacotf&tecnet1,jete,jcs,mil

Joe J Harrison
GTRI

OFF TEL: (904)-862-6229

Donald Hartvigsen
DYNCORP
2727 Hamner
Norco CA 91760
OFF TEL: (909)-735-3300
FAX: (909)-371-7170

Richard E Helmuth
SAIC
Suite 470
8201 Greensboro Drive
McLean VA 22102
OFF TEL: (703)-847-5587
FAX: (703)-847-6406
E-mail: helmuth@tecnet1.jcte.jcs.mil

Capt Eugene H Henry
HQ SMOTEC/OA
606 Cruz Ave
Hurlburt Field FL 32544-5736
OFF TEL: (904)-884-7500 DSN: 579-7500
FAX: (904)-884-5218
E-mail: SMEHENRY@AFSOC.HQAFSOC.AF.MIL

Robert A. Heston
Sverdrup Technology
Suite 4
626 Anchors St., NW
Fort Walton Beach FL 32548
OFF TEL: (904)-833-7600 DSN: 872-6100
FAX: (904)-729-6377
E-mail: AFOTEC.AF.MIL.HESTONRA

Thomas Hilgen
513 ETS/EEN
901 SAC Blvd, Suite 1H1
Offutt AFB NE 68113-5520
OFF TEL: (402)-294-7027 DSN: 271-7027
FAX: (402)-294-6526

Floyd I Hill
IDA
1801 N. Beauregard Street
Alexandria VA 22311-1272
OFF TEL: (703)-845-2053 DSN: 289-1825
FAX: (703)-845-6911

DR Robert G Hinkle
ODUSA(OR)
Attn: SAUS-OR
Pentagon, Room 2D278
Washington DC 20310
OFF TEL: (703)-693-9467 DSN: 227-0367
FAX: (703)-697-7748
E-mail: saaaor0@pentagon-rmh2,army.mil

Joyce A Hires
US Army Test and Evaluation Command
Attn: AMSTE-TA-G
Aberdeen Proving Ground MD 21005-5055
OFF TEL: (410)-278-1434 DSN: 298-1434
FAX: (410)-278-9174
E-mail: jhires@apg-a.apg.army.mil

MajGen Ronald V Hite
OASA(RDA)
103 Army Pentagon
Room 3E448
Washington DC 20310-0103
OFF TEL: (703)-695-3115

Frederick B Hollick Jr.
QUADELTA, Inc.
Suite 302
6201 Leesburg Pike
Falls Church VA 22044-2203
OFF TEL: (703)-237-8990
FAX: (703)-237-9362

Walter W Hollis
DUSA (OR), Hq Dept of the Army
ATTN: SAUS(OR)
Pentagon, Room 2E660
Washington DC 20310-0102
OFF TEL: (703)-695-0083 DSN: 225-0083
FAX: (703)-693-3897

Charles F Horton
Operational T&E (Conventional Systems)
Room 1C730
1700 Defense Pentagon
Washington DC 20301-1700
OFF TEL: (703)-697-3891 DSN: 227-3891
FAX: (703)-614-3992

1LT Thomas J Houle
53 ETS, 901 SAC Blvd Suite 1H1
Offutt AFB NE 68113-5520
OFF TEL: (402)-291-4715 DSN: 271-7027
FAX: (402)-271-6526

RADM William Houley
OPNAV(N091)
Pentagon Room 5C686
Washington DC 20350
OFF TEL: (703)-697-5533

Martha R Hudson
MARCORSYSCOM
Code PSA-O
Quantico VA 22134
OFF TEL: (703)-640-4451 DSN: 278-4451
FAX: (703)-640-2168

Ronald A Jacob
46th Test Wing
Suite 222
101 West D Ave
Eglin AFB FL 32542-5492
OFF TEL: (904)-882-3614 DSN: 872-3614
FAX: (904)-882-9512
E-mail: jacob@ut4.eglin.af.mil

Carroll Jones
HQ USAF/TE
Washington DC 20330

Colleen M Keller
CNA/VX-5
Naval Air Weapons Station
China Lake CA 93555
OFF TEL: (619)-939-4859 DSN: 437-4859
FAX: (619)-939-5744
E-mail: TECNET ID.NO. "CMKELLER"

DR H. Steven Kimmel
BDM Engineering Services Co
1509 BDM Way
McLean VA 22102-3204
OFF TEL: (703)-848-6995
FAX: (703)-848-6990

DR J. Terrence Klopcic
US Army Ballistic Research Laboratory
ATTN: SLCBR-VL-I
Bldg 328
Aberdeen Proving Ground MD 21005-5066
OFF TEL: (410)-278-6322 DSN: 298-6322
FAX: (410)-278-6852
E-mail: klopcic@brl.mil

LT George P Kobler
Commander Operational T&E Force
7970 Diven Street
Norfolk VA 23505
OFF TEL: (804)-444-2954 DSN: 564-2954
FAX: (804)-444-1200
E-mail: peach@TECNET1.jcte.jcs.mil

James C Kolding
Teledyne Brown Engineering
300 Sparkman Dr, NW
PO Box 070007, MS #170
Huntsville AL 35807-7007
OFF TEL: (205)-726-2893
FAX: (205)-726-2241

CDR Grayson L Koogle
Defense Systems Management College
FD-TE
9820 Belvoir Road
Fort Belvoir VA 22060-5565
OFF TEL: (703)-805-2887 DSN: 655-2887
FAX: (703)-805-3183
E-mail: koogle@TECNET1.JCTE.JCS.MIL

Cynthia Kee LaFreniere
Military Operations Research Society
101 S Whiting Street
Suite 202
Alexandria VA 22304
OFF TEL: (703)-751-7290
FAX: (703)-751-8171

David P Leonard
Naval Air Warfare Center
Weapons Div, Code P0391
521 9th Street
Point Mugu CA 93042-5001
OFF TEL: (805)-989-0293 DSN: 351-0293
FAX: (805)-989-0588

DR William G Lese Jr
OASD(PA&E)(GPP/LFD)
The Pentagon, Room 2B256
Washington DC 20301-1800
OFF TEL: (703)-695-0881 DSN: 225-0881
FAX: (703)-693-5707

Janelle Loper
PMO RCAS
8510 Cinderbed Road
Newington VA 22122-8510
OFF TEL: (602)-538-5227
FAX: (602)-538-4336

John F Lyons
JHU/APL
Johns Hopkins Road
Laurel MD 20723
OFF TEL: (301)-412-5201
FAX: (301)-412-1141

LTC Steve G Mankowski
PMO RCAS
2880 LaGrange Circle
Boulder CO 80303-6939
OFF TEL: (303)-497-7335
FAX: (303)-497-5995

DR John W Matherne
US Army Logistics Management College
Decision Sciences Department
Attn: ATSZ-LSS
Fort Lee VA 23801-6050
OFF TEL: (804)-765-4714 DSN: 685-4714
FAX: (804)-734-4164
E-mail: jmathern@alexandria-emh1.army.mil

George W Matrigali
Defense Logistics Agency
Cameron Station
Alexandria VA 22304-6100
OFF TEL: (703)-617-0947

Mason McClanahan
TRADOC
ATTN: ATCD-RM
Fort Monroe VA 23651-5000
OFF TEL: (804)-727-3270 DSN: 680-3270
FAX: (804)-680-2588
E-mail: MCCLANAM@MONROE-EMH1.ARMY.MIL

Gregory K McGill
OSD JTAMS JTF
1431 McGuire Street
Lackland AFB TX 78236-5532
OFF TEL: (210)-671-1910 DSN: 473-1910
FAX: (210)-371-2459

COL James L McKinley
AF/XOME
Evaluation Support Div/MS&A Directorate
1480 Air Force Pentagon
Washington DC 20330-1480
OFF TEL: (202)-507-5336 DSN: 285-5336
FAX: (703)-507-5352
E-mail: lmckin@dmso.dtic.dla.mil

DR Gerald R McNichols
Management Consulting & Research, Inc
5113 Leesburg Pike
Suite 509
Falls Church VA 22041
OFF TEL: (703)-820-4600
FAX: (703)-820-4398

John E Meeuwissen
Georgia Tech Research Institute
Suite 1910
1700 N. Moore St
Arlington VA 22209
OFF TEL: (703)-528-0883
FAX: (703)-528-8419

Capt Kevin D Menard
DET 4 AFOTEC/MIL
4146 E. Bijou Street
Colorado Springs CO 80909-6899
OFF TEL: (719)-554-4086 DSN: 692-4086
FAX: (719)-554-4003

Karen J Merritt
DoD/DESA
2251 Wyoming Blvd, SE
Kirtland AFB NM 87117-5609
OFF TEL: (505)-262-4574
FAX: (505)-262-4621
E-mail: MERRITT@TECNET1.JCTE.JCS.MIL

Richard S Miller
Institute for Defense Analyses
1801 N. Beauregard St
Alexandria VA 22311
OFF TEL: (703)-845-2264 DSN: 289-1868
FAX: (703)-845-6911

Peggy A Mion
Army Test and Evaluation Command
Aberdeen Proving Ground MD 21005-5055
OFF TEL: (410)-278-1441 DSN: 298-1441
FAX: (410)-278-9170
E-mail: pmion@apg-9.army.mil

Maurice Mizrahi
OSD
The Pentagon 2E274,
Washington DC 20301-5707
OFF TEL: (703)-695-5432
FAX: (703)-693-5707

DR Ernest R Montagne
BDM Engineering Svcs. Co.
PO Box 2290
Sierra Vista AZ 85636
OFF TEL: (602)-538-5338 DSN: 879-5338
FAX: (602)-536-4340
E-mail: montagne@cc.sims.disa.mil

William D Moore
USA TEXCOM IEWTD
Fort Huachuca AZ 85613

DR Mark A Moulding
Johns Hopkins University/APL
Johns Hopkins Road
Laurel MD 20723-6099
OFF TEL: (410)-792-6000
FAX: (301)-953-5762
E-mail: markmoulding@JHUapl.edu

Elizabeth C Murter
USACSTA
STECS PO-1E
Aberdeen Proving Ground MD 21005-5059
OFF TEL: (410)-278-9484 DSN: 298-9484
FAX: (410)-278-4116

Sharon R Nichols
HQ AFOTEC/SAN
8500 Gibson Blvd SE
Kirtland AFB NM 87117-5558
OFF TEL: (505)-846-2837 DSN: 246-2837
FAX: (505)-846-5145
E-mail: nichols@hq.afotec.af.mil

Paul A O'Brien
Naval Air Warfare Center Aircraft Div
FW60A
Patuxent River MD 20670-5304
OFF TEL: (301)-826-7592 DSN: 326-7592
FAX: (301)-826-7595

LtCol Phillip O'Brien
Marine Corps Operational T&E Activity
3035 Barnett Ave
Quantico VA 22134
OFF TEL: (703)-640-3141 DSN: 278-3141
FAX: (703)-640-2472

James F O'Bryon
OUSD (A)
DDDRE(T&E)ODLFT
Pentagon, Room 3E1060
Washington DC 20301-3110
OFF TEL: (703)-697-5732 DSN: 227-5732
FAX: (703)-614-9883

RADM Daniel T Oliver
Office Chief of Naval Operations (N81)
Assessments Division, Room 4A530
2000 Navy Pentagon
Washington DC 20350-2000
OFF TEL: (703)-697-0831 DSN: 227-0831
FAX: (703)-695-6903

William Owen
U.S. Army Medical Department Board
US Army Medical Dept Center & School
ATTN: HSMC-FBS, 1961 Wilson Street
Fort Sam Houston TX 78234-6124
OFF TEL: (210)-221-2219 DSN: 471-2219
FAX: (210)-554-4777

Richard W Pace
OUSD(A&T)/DT&E
Pentagon, Room 3D1067
Washington DC 20301-3110
OFF TEL: (000)-070-3697 DSN: 227-4818
FAX: (703)-614-9103

Paul D Parker
Vector Data Systems
Suite 300
1100 S. Washington Street
Alexandria VA 22314-4400
OFF TEL: (703)-418-0771
FAX: (703)-418-0774

Harold C Pasini
US Army OPTEC
Park Center IV
4501 Ford Avenue
Alexandria VA 22302-1458
OFF TEL: (703)-756-2294
FAX: (703)-746-0498
E-mail: pasini%tex3@texcom.emh1.army.mil

Kathryn P Pearson

OFF TEL: (708)-491-2795
FAX: (708)-491-8005

Richard E Pearsall
PRC, Inc.
468 Viking Drive
Virginia Beach VA 23452
OFF TEL: (804)-444-6153 DSN: 564-6153
FAX: (804)-444-1880

Charles O Pflugrath
BDM Engineering Services Co
1509 BDM Way
McLean VA 22102
OFF TEL: (703)-848-5810
FAX: (703)-848-5216

R Alan Plishker
International Test and Evauluation Assn
4400 Fair Lakes Court
Fairfax VA 22033-3899
OFF TEL: (703)-631-6220
FAX: (703)-631-6221

Raymond G Pollard III
TECOM
USA Test & Evaluation Command
Attn: AMSTE-TD
Aberdeen Proving Ground MD 21005-5055
OFF TEL: (410)-278-1016 DSN: 298-1016
FAX: (410)-278-9029
E-mail: amstetd@apg-9.apg.army.mil

LCDR James C Proulx
Commander Operational T&E Force
7970 Diven Street
Norfolk VA 23505-1498
OFF TEL: (804)-444-2954 DSN: 564-2954
FAX: (804)-565-8516

John R Quinn
Naval Ordnance Center - PAC Div
Human Resources Ofc, Code 063
800 Seal Beach Blvd
Seal Beach CA 90740-5000
OFF TEL: (310)-594-7676 DSN: 873-7676
FAX: (310)-594-7200

Richard A Rabel
Naval Undersea Warfare Center Division
610 Dowell Street
Keyport WA 98345-7610
OFF TEL: (206)-396-2667 DSN: 744-2667
FAX: (206)-396-5189

MAJ Matthew P Ransdell
USAQMC&S
Fort Lee VA 23801
OFF TEL: (804)-734-6360 DSN: 687-6360
FAX: (804)-734-6848

Joseph A Rech
AF/TER
Room 4D866
1650 AF Pentagon
Washington DC 20330-1650
OFF TEL: (703)-693-6597 DSN: 223-6597
FAX: (000)-695-5124
E-mail: Rechj@TECNET1.TCTE.JCS.MIL

LtCol Stalker E Reed
AFOTEC/TEZ
8500 Gibson Blvd SE
Kirtland AFB NM 87117-5558
OFF TEL: (505)-846-8450 DSN: 246-8450
FAX: (505)-846-5214
E-mail: REEDSE@HQ.AFOTEC.AF.MIL@WINS

John A Riente
HQ Department of the Army
Deputy Chief of Staff for Ops & Plans
400 Army Pentagon
Washington DC 20310-0400
OFF TEL: (703)-697-4113 DSN: 227-4113
FAX: (703)-614-9044
E-mail: riente@pentemh2.army.mil

DR William J Riley
Logicon RDA
105 E. Vermijo
Suite 450
Colorado Springs CO 80919
OFF TEL: (719)-635-2571

MajGen Robert Rosenkranz
US Army OPTEC
Attn: CSTE-EZ
4501 Ford Ave
Alexandria VA 22302-1458

DR Carl T Russell
US Army TEXCOM Experimentation Center
ATTN: CSTE-TEC-T
Fort Hunter Liggett CA 93928-5000
OFF TEL: (408)-385-2728 DSN: 359-2728
FAX: (408)-385-1190
E-mail: russellc%ote2@leav-emh.army.mil

DR Patricia A Sanders
OSD(PA&E)
Land Forces
Pentagon, Room 2B256
Washington DC 20301-1800
OFF TEL: (703)-697-3521 DSN: 227-3521
FAX: (703)-693-5707
E-mail: psanders@dmso.dtic.dla.mil

Steven B Schorr
USATRAC-SAC
Attn: ATRC-SAS
Fort Leavenworth KS 66027-5200
OFF TEL: (913)-684-7386 DSN: 552-7386
FAX: (913)-684-3866
E-mail: schorrs@tracer.army.mil

Capt Garry S Schwartz
Marine Corps Operational T&E Acty
3035 Barnett Ave
Quantico VA 22134-5014
OFF TEL: (703)-640-3141 DSN: 278-3286
FAX: (703)-640-2472
E-mail: GIDM2F:MQGMCOTE OPERATIONS ANALYST 2

James B. Sebolka
Science Applications Internatioinal
8201 Greensboro Drive, Suite 470
McLean VA 22102
OFF TEL: (703)-847-5586
FAX: (703)-847-6406

DR Ernest A. Seglie
DOT&E/OSD
Pentagon, Room 3E318
1700 Defense Pentagon
Washington DC 20301-1700
OFF TEL: (703)-697-7247 DSN: 227-7247
FAX: (703)-693-5248
E-mail: seglie-ernest@mail-host.dote.osd.mi

William D Sieg
QUADELTA, Inc
Suite 302
6201 Leesburg Pike
Falls Church VA 22044-2203
OFF TEL: (703)-237-8990
FAX: (703)-237-9362

Anthony W Sinden
British Defence Staff
British Embassy
3100 Massachutetts Avenue
Washington DC 20008
OFF TEL: (202)-898-4288
FAX: (202)-898-4581

Philip C Smith
Johns Hopkins Univ/APL
Johns Hopkins Road
Laurel MD 20723-6099
OFF TEL: (301)-953-5156
FAX: (301)-953-5762
E-mail: smithpcl@central.ssd.jhuapl.edu

Capt Christopher B Snyder
Marine Corps Systems Command
Code PSE-T
2033 Barnett Ave, #315
Quantico VA 22134
OFF TEL: (703)-640-5963 DSN: 278-5963
FAX: (703)-640-3432
E-mail: SNYDER@PSE@MARCORSYSCOM

MAJ Scott E Shaw
Marine Corps Air Facility
2101 Rowell Road
Quantico VA 22134-5064
OFF TEL: (703)-640-3303 DSN: 278-3303
FAX: (703)-640-2224

John R Sinclair
US Coast Guard (G-AT)
2100 2nd Street, SW
Washington DC 20593-0001
OFF TEL: (202)-267-1127
FAX: (202)-267-4279
E-mail: jack_sinclair/g-a@cgsmtp.comdt.uscg

Kevin C Smith
Commander Operational T&E Force
7970 Diven Street
Norfolk VA 23505
OFF TEL: (804)-444-2954 DSN: 564-2954
FAX: (804)-444-1200
E-mail: kcsmith@tecnet1.jcte.jcs.mil

William L Smith
PM, FARV
SMCAR-SS-DF, Bldg 3159
Picatinny Arsenal NJ 07806-5000
OFF TEL: (201)-724-7621 DSN: 880-7621
FAX: (201)-724-7606

Dean Spencer
Scientific Research Corp.
280 Interstate N Parkway Suite 430
Atlanta GA 30339
OFF TEL: (404)-859-9161

E-mail: WDS@Scires.Com

DR Duane Steffey
National Research Council
National Academy of Sciences
2101 Constitution Ave, NW
Washington DC 20418
OFF TEL: (202)-334-1932
FAX: (202)-334-3751

Arthur Stein FS
Institute for Defense Analyses
Operations Evaluation Div
1801 N. Beauregard St
Alexandria VA 22311
OFF TEL: (703)-845-6980
FAX: (703)-845-2588

Robert L Stovall
460G/OGML
104 Cherokee Avenue
Eglin AFB FL 32542-5600
OFF TEL: (904)-882-9243

James J Streilein
USA Material Sys Analysis Act
Attn: AMXSY-R
Aberdeen Proving Ground MD 21005-5071
OFF TEL: (410)-278-6580 DSN: 298-6580
FAX: (410)-278-6584
E-mail: strln@amsaa-cleo.brl.mil

RADM George H Strohsahl
Pacific Missile Test Center
Naval Air Warfare Center
Point Mugu CA 93042

Frank Strong
Department of the Army
PEO-Field Artillery Systems, B 171
Picatinny Arsenal NJ 07806-5000
OFF TEL: (201)-724-7104 DSN: 880-7124

Jacqueline K Telford
Johns Hopkins Univ/APL
Johns Hopkins Road
Laurel MD 20723
OFF TEL: (301)-953-5000
FAX: (301)-953-5762

LtCol Charles H Thomas Jr.
AFSAA/SAG
Pentagon, Room 1D380
Washington DC 20330-1570
OFF TEL: (703)-695-5282 DSN: 225-5282
FAX: (703)-697-3441

Clayton J Thomas FS
AFSAA/SAN
1570 Air Force Pentagon
Room 1E386
Washington DC 20330-1570
OFF TEL: (703)-697-4300 DSN: 227-4300
FAX: (703)-697-3441
E-mail: thomasc@afsaa.hq.af.mil

Lane Thompson
46 Test Wing
TSRR
Eglin AFB FL 32542
OFF TEL: (904)-882-2528

DR Lowell H Tonnessen
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria VA 22311
OFF TEL: (703)-845-6921
FAX: (703)-845-6911

MajGen Richard W Tragemann
US Army TECOM
Aberdeen Proving Ground MD 21005-5055
OFF TEL: (410)-278-1003
FAX: (410)-278-9029

Dyrck H Van Dusen
The SURVICE Engineering Co
Suite 103
1003 Old Philadelphia Rd
Aberdeen MD 21001
OFF TEL: (410)-273-7722
FAX: (410)-272-7417

Eugene P Visco FS
US Army MISMA
Crystal Square 2, #808, SFUS-MIS
1725 Jefferson Davis Hwy
Arlington VA 22202
OFF TEL: (703)-607-3420 DSN: 327-3420
FAX: (703)-607-3381
E-mail: visco@pentagon-hqdadss.army.mil

Michele D Wagner
Naval Air Warfare Center
Weapons Div, Code P0391
521 9th Street
Point Mugu CA 93042-5001
OFF TEL: (805)-989-0409 DSN: 351-0409
FAX: (805)-989-0588

Glenn H Waldron
Defense Evaluation Support Activity
Suite 503
5201 Leesburg Pike
Falls Church VA 22041
OFF TEL: (703)-931-8104
FAX: (703)-931-3663

George G Wauer
OSD/DOT&E
1700 Defense Pentagon
Washington DC 20301-1700
OFF TEL: (703)-614-2520 DSN: 224-2520
FAX: (703)-614-3992

Larry L West
US Army Test & Evaluation Command
Attn: AMSTE-TA-G
Aberdeen Proving Ground MD 21005-5055
OFF TEL: (410)-278-1348 DSN: 298-1348
FAX: (301)-298-9174
E-mail: lwest@apg-9.apg.army.mil

Ralph F Wetzl
Science Applications Internat'l Corp
Suite 470
8201 Greensboro Drive
McLean VA 22102
OFF TEL: (703)-847-5575
FAX: (703)-734-8318

Robert R Wilbur
HQ AFOTEC/TSP
8500 Gibson Blvd
Kirtland AFB NM 87117-5558
OFF TEL: (505)-846-4206 DSN: 246-4206
FAX: (505)-846-5236

John G Wilcox
SRI International
1611 N. Kent Street
Arlington VA 22209
OFF TEL: (703)-247-8467
FAX: (703)-527-3087
E-mail: Wilcox@CRVAX.SRI.COM

Richard I Wiles
Military Operations Research Society
101 S Whiting Street
Suite 202
Alexandria VA 22304
OFF TEL: (703)-751-7290
FAX: (703)-751-8171
E-mail: rwiles@dgis.dtic.dla.mil

George G Williams
PEO TAC MSLS
Attn: SFAE-MSL
Redstone Arsenal AL 35898-8000
OFF TEL: (205)-876-0714 DSN: 746-0714
FAX: (205)-955-9443

DR Marion L Williams FS
HQ AFOTEC/CN
8500 Gibson Blvd SE
Kirtland AFB NM 87117-5558
OFF TEL: (505)-846-0607 DSN: 246-0607
FAX: (505)-846-9726
E-mail: wiliamsm@hq.afotec.af.mil

COL Stephen D Williams
US Army War College
Center for Strategic Leadership
Carlisle PA 17013-5050
OFF TEL: (717)-245-3165 DSN: 242-3165
FAX: (717)-245-3279

Phillip E Wralstad
IEWTD
Hayes Hall
Fort Huachuca AZ 85613
OFF TEL: (602)-538-8814 DSN: 879-8814
FAX: (602)-538-8815

Alice W Yee
US Army Total Cost & Econ Analysis Ctr
5611 Columbia Pike
Falls Church VA 22041-5050
OFF TEL: (703)-756-2018 DSN: 289-2018
FAX: (703)-756-7553

RADM John J Zerr
Commander Operational T&E Force
7970 Diven Street
Norfolk VA 23505-1498
OFF TEL: (804)-444-5162 DSN: 564-5162
FAX: (804)-445-8932